Geometric optimization problems in quantum computation and discrete mathematics: Stabilizer states and lattices

Inaugural-Dissertation
zur
Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln
vorgelegt von

Arne Heimendahl
aus Krefeld

2023

Berichterstatter (Gutachter):	Prof. Dr. Frank Vallentin Prof. Dr. Michael Walter
Tag der mündlichen Prüfung:	21.02.2023

Abstract

This thesis consists of two parts:

Part I deals with properties of stabilizer states and their convex hull, the stabilizer polytope. Stabilizer states, Pauli measurements and Clifford unitaries are the three building blocks of the stabilizer formalism whose computational power is limited by the Gottesman-Knill theorem. This model is usually enriched by a magic state to get a universal model for quantum computation, referred to as quantum computation with magic states (QCM). The first part of this thesis will investigate the role of stabilizer states within QCM from three different angles.

The first considered quantity is the stabilizer extent, which provides a tool to measure the non-stabilizerness or magic of a quantum state. It assigns a quantity to each state roughly measuring how many stabilizer states are required to approximate the state. It has been shown that the extent is multiplicative under taking tensor products when the considered state is a product state whose components are composed of maximally three qubits. In Chapter 2, we will prove that this property does not hold in general, more precisely, that the stabilizer extent is strictly submultiplicative. We obtain this result as a consequence of rather general properties of stabilizer states. Informally our result implies that one should not expect a dictionary to be multiplicative under taking tensor products whenever the dictionary size grows subexponentially in the dimension.

In Chapter 3, we consider QCM from a resource theoretic perspective. The resource theory of magic is based on two types of quantum channels, completely stabilizer preserving maps and stabilizer operations. Both classes have the property that they cannot generate additional magic resources. We will show that these two classes of quantum channels do not coincide, specifically, that stabilizer operations are a strict subset of the set of completely stabilizer preserving channels. This might have the consequence that certain tasks which are usually

realized by stabilizer operations could in principle be performed better by completely stabilizer preserving maps.

In Chapter 4, the last one of Part I, we consider QCM via the polar dual stabilizer polytope (also called the Λ-polytope). This polytope is a superset of the quantum state space and every quantum state can be written as a convex combination of its vertices. A way to classically simulate quantum computing with magic states is based on simulating Pauli measurements and Clifford unitaries on the vertices of the Λ-polytope. The complexity of classical simulation with respect to the polytope Λ is determined by classically simulating the updates of vertices under Clifford unitaries and Pauli measurements. However, a complete description of this polytope as a convex hull of its vertices is only known in low dimensions (for up to two qubits or one qudit when odd dimensional systems are considered). We make progress on this question by characterizing a certain class of operators that live on the boundary of the Λ-polytope when the underlying dimension is an odd prime. This class encompasses for instance Wigner operators, which have been shown to be vertices of Λ. We conjecture that this class contains even more vertices of Λ. Eventually, we will shortly sketch why applying Clifford unitaries and Pauli measurements to this class of operators can be efficiently classically simulated.

Part II of this thesis deals with lattices. Lattices are discrete subgroups of the Euclidean space. They occur in various different areas of mathematics, physics and computer science. We will investigate two types of optimization problems related to lattices.

In Chapter 6 we are concerned with optimization within the space of lattices. That is, we want to compare the Gaussian potential energy of different lattices. To make the energy of lattices comparable we focus on lattices with point density one. In particular, we focus on even unimodular lattices and show that, up to dimension 24, they are all critical for the Gaussian potential energy. Furthermore, we find that all n-dimensional even unimodular lattices with $n \leq 24$ are local minima or saddle points. In contrast in dimension 32, there are even unimodular lattices which are local maxima and others which are not

even critical.

In Chapter 7 we consider flat tori $\mathbb{R}^n/L$, where L is an n-dimensional lattice. A flat torus comes with a metric and our goal is to approximate this metric with a Hilbert space metric. To achieve this, we derive an infinite-dimensional semidefinite optimization program that computes the least distortion embedding of the metric space $\mathbb{R}^n/L$ into a Hilbert space. This program allows us to make several interesting statements about the nature of least distortion embeddings of flat tori. In particular, we give a simple proof for a lower bound which gives a constant factor improvement over the previously best lower bound on the minimal distortion of an embedding of an n-dimensional flat torus. Furthermore, we show that there is always an optimal embedding into a finite-dimensional Hilbert space. Finally, we construct optimal least distortion embeddings for the standard torus $\mathbb{R}^n/\mathbb{Z}^n$ and all 2-dimensional flat tori.

Contents

Part I

The Stabilizer World

Chapter 1

Introduction - Stabilizer states and measures of magic

The *stabilizer formalism* is a widely-used theoretical framework in quantum computation. It constitutes one of the cornerstones to build fault-tolerant quantum computers. The three main objects of the stabilizer formalism are *stabilizer states, Clifford unitaries* and *Pauli measurements*. However, quantum systems that are only built upon these three primitives, also called stabilizer circuits, can be simulated on a classical computer in polynomial time in the number of qubits – this is the content of the famous Gottesman-Knill theorem [Got99, SA04].

To promote this model to a universal model for a quantum computer, one requires an extra non-classical resource, commonly called *magic*. Usually, magic stems from a so called "magic state" which is "injected" into the stabilizer circuit [BK05]. A necessary condition for a state to possess magic is that it is not contained in the convex hull of stabilizer states, the *stabilizer polytope*. When a stabilizer circuit is enriched by a magic state, we call this *quantum computing with magic states* (QCM), which is a universal model for quantum computation [BK05].

In recent years, a lot of research has been devoted to understand the nature of magic in quantum states. A particularly prominent way to measure magic is by means of classical simulation cost. Here, some quantity, often called the *magic monotone*, is assigned to the magic input state $|\psi\rangle$. The magic monotone governs the cost of classically simulating QCM with input state $|\psi\rangle$.

Classical simulation methods can be roughly divided into two classes: *quasiprobability methods* [ME12, VFGE12, VMGE14, PWB15, BVDB$^+$17, HC17a, FRB18, RBVT$^+$20, HG19, PRSKB22] and *stabilizer rank methods* [BSS16, SC19, SRP$^+$21, BG16, QPG21, PSV22, BGL22, Koc20]. In this thesis, we will not study classical simulation methods for QCM but we will rather take a closer look at some of the quantities and objects that these ideas are based on.

Chapter 2: Stabilizer rank and the stabilizer extent

A particular fruitful way to design classical simulation algorithms for QCM is based on the *stabilizer rank* [BSS16, BG16]. Let STAB_n be the set of n-qubit stabilizer states. Given an n-qubit state $|\psi\rangle$, the stabilizer rank is given by

$$\chi(\psi) = \min\left\{R \,:\, |\psi\rangle = \sum_{i=1}^{R} c_i \,|s_i\rangle\,,\; c_i \in \mathbb{C},\; |s_i\rangle \in \mathrm{STAB}_n\right\}.$$

If a classical simulation algorithm for QCM with input state $|\psi\rangle$ relies on a decomposition of the above from, then its runtime is typically governed by $\chi(\psi)$.

As an ℓ_0-minimization problem, explicitly computing the stabilizer rank is NP-hard, see [FR13, Section 2.3], even for particular classes of states. Non-trivial[1] exponential upper bounds [BSS16, QPG21, Koc20] and linear lower bounds [PSV22, Lab22, LS22] have been constructed for the stabilizer rank of the most promising candidates for magic states. Currently, it is an open problem to close this huge gap. In general, it is widely believed that magic states have an exponential stabilizer rank. The reason behind this is that QCM with n qubits can be classically simulated in $\chi(H^{\otimes n})^2 n^4$ [SA04],[BSS16, p. 3], where $|H\rangle = \cos(\pi/8)\,|0\rangle + \sin(\pi/8)\,|1\rangle$ is a particular magic state. Now, if $\chi(H^{\otimes n})^2$ was polynomial in n, there would be a (randomized) polynomial time classical algorithm for QCM. In complexity theoretic terms this would imply BPP = BQP, which is not believed to be true.

[1] This means that the bounds are better than the trivial upper bound of 2^n.

A numerically more approachable quantity is the *stabilizer extent* [BBC+19], which is the ℓ_1-relaxation of the stabilizer extent: For some n-qubit state $|\psi\rangle$, it is given by

$$\xi(\psi) = \min\left\{ \Big(\sum_{s\in \mathrm{STAB}_n} |c_s| \Big)^2 \;:\; |\psi\rangle = \sum_{s\in \mathrm{STAB}_n} c_s \, |s\rangle \,,\, c_s \in \mathbb{C} \right\}.$$

In contrast to the stabilizer rank, computing the extent can be expressed as a second order cone problem and is solvable in polynomial time in the input size [AG03]. However, as the input size is still exponential in the number of qubits, numerically computing the extent is only possible for a few qubits.

Instead of computing χ or ξ explicitly, oftentimes it suffices to compute a "good" stabilizer decomposition of the input state $|\psi\rangle$, which gives an upper bound for χ or ξ. In fact, any decomposition of $|\psi\rangle$ as a linear combination of stabilizer states can be fed into a classical simulation algorithm for QCM which is based on the stabilizer rank or extent. To compute upper bounds one typically exploits that both quantities are sub-multiplicative under taking tensor products, that is

$$\chi(\psi \otimes \phi) \leq \chi(\psi)\chi(\phi) \quad \text{and} \quad \xi(\psi \otimes \phi) \leq \xi(\psi)\xi(\phi)$$

for any two states ψ and ϕ. Surprisingly, if ψ and ϕ are just composed of up to three qubits, then the above inequality is tight for the extent [BBC+19], i.e. $\xi(\psi \otimes \phi) = \xi(\psi)\xi(\phi)$. Consequently, computing the extent for product states $|\psi_1\rangle \otimes \cdots \otimes |\psi_m\rangle$, where each component is composed of maximally three qubits, boils down to solving m second order cone problems with small input size.

It was an open question whether this property holds in full generality, i.e. for product states where each component has an arbitrary number of qubits. In Chapter 2 we will provide a negative answer to this question. More precisely, we show that if n is sufficiently large and $|\psi\rangle$ is an n-qubit Haar-randomly chosen state, then with high probability

$$\xi(\psi \otimes \psi^*) < \xi(\psi)\xi(\psi^*).$$

To prove the result, we make use of the dual formulation of ℓ_1-minimization problems with complex coefficients. Therefore, let $\mathcal{D} \subset \mathbb{C}^d$ be some dictionary and define the extent with respect to $\mathcal{D}$ by

$$\xi_{\mathcal{D}}(x) = \min \left\{ \left(\sum_{s \in \mathcal{D}} |c_s| \right)^2 \, : \, c \in \mathbb{C}^{|\mathcal{D}|}, \, x = \sum_{s \in \mathcal{D}} c_s s \right\}.$$

Writing this as a second order cone problem and then dualizing gives the following dual formulation:

$$\begin{array}{rcll} \xi_{\mathcal{D}}(x) & = & \max & |\langle x, y \rangle|^2 \\ & & \text{s.t.} & y \in \mathbb{C}^d, \\ & & & \max_{s \in \mathcal{D}} |\langle y, s \rangle|^2 \leq 1. \end{array} \tag{1.1}$$

Using the primal and the dual formulation, one can show that $\xi_{\mathcal{D} \otimes \mathcal{D}}(x \otimes x') = \xi_{\mathcal{D}}(x) \xi_{\mathcal{D}}(x')$ for all $x, x' \in \mathbb{C}^d$, where $\mathcal{D} \otimes \mathcal{D}$ refers to the product dictionary $\mathcal{D} \otimes \mathcal{D} = \{ s \otimes s' \, : \, s, s' \in \mathcal{D} \}$. The optimal dual solution of (1.1) for $\xi_{\mathcal{D} \otimes \mathcal{D}}(x \otimes x')$ is generically unique and of the form $y \otimes y'$. Now, to prove that

$$\xi_{\tilde{\mathcal{D}}}(x \otimes x') < \xi_{\mathcal{D} \otimes \mathcal{D}}(x \otimes x')$$

for some dictionary $\tilde{D}$ which strictly contains $\mathcal{D} \otimes \mathcal{D}$, it suffices to show that there is $\tilde{s} \in \tilde{D} \setminus (\mathcal{D} \otimes \mathcal{D})$ such that $|\langle \tilde{s}, y \otimes y' \rangle|^2 > 1$, implying that $y \otimes y'$ is infeasible for the dual formulation of $\xi_{\tilde{\mathcal{D}}}$. In the case case of stabilizer states, one can show that generically

$$\xi_{\mathrm{STAB}_{2n}}(\psi \otimes \psi^*) \leq \xi_{(\mathrm{STAB}_n \otimes \mathrm{STAB}_n) \cup \{\Phi\}}(\psi \otimes \psi^*) < \xi_{\mathrm{STAB}_n \otimes \mathrm{STAB}_n}(\psi \otimes \psi^*)$$

for a Haar-random state ψ and $\Phi \in \mathrm{STAB}_{2n}$ being the maximally entangled state.

In summary, computing the stabilizer extent of general product states remains a hard problem. For the same reason, if $|\psi_1\rangle \otimes \cdots \otimes |\psi_m\rangle$ is a general product state, then a stabilizer decomposition that is obtained by optimally decomposing the components $|\psi_i\rangle$ into stabilizer states with respect to the ℓ_1-norm, is in general not the best input for stabilizer extent based classical simulation methods of QCM.

Our proof shows that the strict sub-multiplicativity of the extent follows from rather general properties of stabilizer states. Among

those, the most important property is that the number of stabilizer states grows subexponentially in the underlying dimension (in fact quadratically, as there are $\mathcal{O}(N^2)$ stabilizer states in dimension $N = 2^n$) [Gro06, Corollary 21]. Hence, the result can be seen as an indicator that minimizing the ℓ_1-norm with respect to general dictionaries is not multiplicative under taking tensor products – provided that the size of the dictionary grows subexponentially in the dimension.

Chapter 3: The resource theory of magic

Another way to understand the phenomena that make QCM non-classical comes from a more abstract point of view – the perspective of *quantum resource theories* (for a survey about this topic see [CG19]).

In quantum resource theories, the set of quantum states is usually partitioned into *free states* and *resource states*. Vaguely speaking, the latter is supposed to represent the quantumness of the considered model.

For our setting in QCM, we choose the set of free states to be the stabilizer polytope. The exact form of this partition is a matter of discussion – we could also choose the set of free states to be those that live in the convex hull of cnc-operators as defined in [RBVT$^+$20] (cnc refers to closed under inference and non-contextual), or, in the case of odd prime dimensional systems, those that have a non-negative Wigner function [Gro06]. Intuitively, the requirement is that the set of free states is somewhat considered as classical within the resource theory of interest.

A resource theory comes with a set of *free operations*. These are operations on the set of quantum states which preserve the set of free states and which are *resource non-generating*. This means that any quantity that assigns some value to the amount of resource of a quantum state (for example the stabilizer rank or extent) is supposed to be non-increasing under applying free operations.

Usually, there are two ways to define free operations. The first one is *operational*. This means that there is an explicit set of fundamental free operations that can be composed in some way to obtain more

free operations – these can be seen as the generators of the set of free operations.

The second one is *axiomatically*. In this case, we simply define the set of free operations to be the ones that preserve the set of free states.

In QCM and the operational setup, free operations are *stabilizer operations* (SO). Here the fundamental free operations are preparing stabilizer states, applying Clifford unitaries and doing Pauli measurements. To build a general stabilizer operation, we are allowed to apply the three operations in any order and as often as we want, and we are allowed to apply them according to a probability distribution of our choice.

In contrast, from an axiomatic point of view, free operations are precisely those that preserve the stabilizer polytope, referred to as *completely stabilizer preserving maps* [SC19] (CSP).

One can easily verify that every stabilizer operation is also completely stabilizer preserving, i.e. SO $\subseteq$ CSP. However, it was an open question whether this inclusion is strict or whether both definitions lead to the same class of operations.

This question has a famous counterpart in entanglement theory. In this setup, the set of free states is the set of separable states and resource states are entangled states. Operationally free refers to local operations and classical communication (LOCC), whereas, from an axiomatic point of view, quantum channels that preserve the set of separable states constitute the set of free operations. It was proven that LOCC is a strict subset of the set of separable maps [BDF$^+$99, CLM$^+$14].

In Chapter 3, we will show that this inclusion is also strict for the resource theory of magic, that is SO $\subsetneq$ CSP. We achieve this by explicitly constructing a quantum channel which is completely stabilizer preserving but not a stabilizer operation. This quantum channel maps n qudits to n qudits where $n \geq 2$ and the underlying qudit dimension can be 2 or an odd prime. Furthermore, we show that stabilizer operations and completely stabilizer preserving maps coincide in the regime of channels that map one qudit to one qudit.

Our arguments are roughly built upon the following ideas. Via the

Choi-Jamiołkowski-isomorphism it is not hard to see that the set of CSP-maps sending n qudits to n qudits, CSP_n, is a polytope [SC19]. Using this characterization, one can show that every CSP-map can be written as [Hei21]

$$\mathcal{E}(\rho) = \sum_{i=1}^{r} \lambda_i \frac{d^n}{\operatorname{rank} P_i} U_i P_i \rho P_i U_i^\dagger. \tag{1.2}$$

Here, d is the underlying qudit dimension, the P_i's are stabilizer code projectors onto stabilizer codes, the U_i's are Clifford unitaries and the λ_i form a probability distribution. Since CSP-maps are quantum channels and thus trace preserving we have the extra condition:

$$\mathbb{1} = \mathcal{E}^\dagger(\mathbb{1}) = \sum_{i=1}^{r} \lambda_i \frac{d^n}{\operatorname{rank} P_i} P_i. \tag{1.3}$$

The set SO has a more intricate structure and we do not know whether it is a polytope (however, this seems very plausible, as will be explained in the sequel). Analogously to (1.2), there is also a normal form for the set of SO-maps, sending n to n qudits, SO_n:

$$\mathcal{E}(\rho) = \mathrm{Tr}_{n+1,\ldots,n+r} \sum_i U_i \left(P_i \rho P_i \otimes |0^r\rangle\langle 0^r|\right) U_i^\dagger, \tag{1.4}$$

where $\{P_i\}$ is a projective measurement given by mutually orthogonal stabilizer code projectors which sum up to the identity, and the U_i's are Clifford unitaries acting on $n + r$ qudits. Every stabilizer operation can be written as a convex combination of operations of the above form.

Comparing these two normal forms, the crucial difference is that the P_i's in (1.4) need to be *orthogonal* projectors whereas in (1.2), they only need to satisfy (1.3) (this is also satisfied for SO because the P_i's sum up to the identity). This gives rise to our strategy to build a channel that is CSP but not SO: simply construct an (extremal[2]) CSP-channel of the form (1.2) where the corresponding projectors are not mutually orthogonal.

In principle, the number of ancilla qubits r in (1.4) need not to be bounded. However, if it was possible to bound the number of required

[2]This means that the channel is a vertex of the polytope of CSP-channels.

ancialla qubits, this would show that SO_n is a polytope. To see this, it suffices to note that there is only a finite number of channels of the form (1.4) for fixed n and r.

Comparing this to the resource theory of entanglement, we observe differences in the geometry of the underlying sets. The set of quantum channels that preserve the set of separable states is not a polytope and set of LOCC operations is not even closed [CLM+14]. Arguably, the resource theory of magic can be seen a discrete counterpart of the resource theory of entanglement.

Chapter 4: Quasiprobability methods and the Λ-polytope

For classical simulation of QCM based on quasiprobability methods, the magic input (and possibly mixed) state ρ is typically expressed as an $\mathbb{R}$-linear combination of some generating set $\mathcal{K} \subset \mathrm{Herm}_1(d^n)$, where $\mathrm{Herm}_1(d^n)$ is the set of $d^n \times d^n$ Hermitian matrices of trace one:

$$\rho = \sum_{M \in \mathcal{K}} c_M M, \quad c_M \in \mathbb{R}.$$

For example the set $\mathcal{K}$ could be the set of stabilizer states or, in odd dimensional systems, the set of Wigner operators. Using Monte-Carlo sampling techniques, one can design classical simulation algorithms for QCM whose complexity is determined by one of the two relevant quantities:

1. The ℓ_1-norm of the expansion coefficients [VFGE12, PWB15, HC17b], which intuitively measures how much $\{c_M\}_{M \in \mathcal{K}}$ deviates from a probability measure. Due to, $\mathrm{Tr}(M) = 1$ for all $M \in \mathcal{K}$, it is determined by

$$\sum_{M \in \mathcal{K}} |c_M| = \sum_{c_M > 0} c_M - \sum_{c_M < 0} c_M = 1 + 2 \sum_{c_M < 0} |c_M|.$$

2. The complexity of classically simulating the updates of elements $M \in \mathcal{K}$ under Pauli measurements and Clifford unitaries [Zur20].

If $\mathcal{K}$ is the set of stabilizer state projectors [HC17a, HG19] or the set of cnc-operators classified in [RBVT+20], or the set of phase point

operators that define the Wigner function [VFGE12, VMGE14] (for odd-dimensional systems), then simulation tasks that fall under 2. (classically simulating Pauli measurements and Clifford unitaries) can be efficiently realized [3]. In these cases, the runtime of a classical simulation algorithm for QCM is governed by the amount of negativity in the coefficients $\sum_{c_m<0} |c_M|$. The corresponding classical simulation methods belong to the class of quasiprobability methods.

In contrast, one may also choose the set $\mathcal{K}$ to be the set of vertices of the *polar dual stabilizer polytope* [4] (also referred to as the Λ-polytope) [Hei19, ZOR20]. Let STAB_n be the set of n-qudit stabilizer states, viewed as rank-1 density matrices. Then Λ-polytope is given by

$$\Lambda_n = \{X \in \mathrm{Herm}_1(d^n) \,:\, \mathrm{Tr}(SX) \geq 0 \;\text{ for all } S \in \mathrm{STAB}_n\}.$$

This set is indeed a polytope and every quantum state can be written as a convex combination of the vertices of Λ_n:

$$\rho = \sum_{M \in \mathrm{Vert}(\Lambda_n)} c_M M \quad \text{with} \quad c_M \geq 0, \sum_{M \in \mathrm{Vert}(\Lambda_n)} c_M = 1.$$

Moreover, Pauli measurements and Clifford unitaries preserve the Λ-polytope, in the sense that if an operator is contained in the Λ-polytope, then also its image after applying Pauli measurements or Clifford unitaries (in particular, this holds for the vertices of Λ).

This allows us to classically simulate QCM in the following way: Sample an operator M according to the probability distribution $\{c_M\}$, then compute the evolution of M under Clifford unitaries and Pauli measurements. The runtime of this algorithm is determined by classically performing the two tasks: Sampling from $\{c_M\}$ and computing the updates of elements in $\mathcal{K}$ under Clifford unitaries and Pauli measurements. Unfortunately, this approach bears a major drawback – a complete list of vertices of Λ is only known for a very few cases [Hei19, CGG$^+$06, Rei05, OZR21] and it seems to be an extremely hard task to classify all of them.

[3] This means in polynomial time in the number qubits/qudits.

[4] In fact, we extend the standard notion of polar dual polytope to "polar dual polytope contained in an affine subspace"; see [ZORH21, Appendix C] for an explanation.

In Chapter 4, we nevertheless try to make some progress on this question. We consider odd-prime dimensional systems and focus on a class of operators in Λ_n among which some have already been shown to be vertices. These operators have the property that when they are expanded in the generalized Pauli basis, then their expansion coefficients are roots of unity. This means we want to characterize $A_\Omega^\eta \in \Lambda_n$ with

$$A_\Omega^\eta = \frac{1}{d^n} \sum_{u \in \Omega} \omega^{\eta(u)} T(u), \qquad \Omega \subset \mathbb{F}_d^{2n},\ \eta : \Omega \to \mathbb{R}, \tag{1.5}$$

where $\omega = e^{2\pi i/d}$ and $T(u)$ are the elements of the generalized Pauli basis, labeled by points $u \in \mathbb{F}_d^{2n}$. This class of operators encompasses stabilizer states (Ω is a Lagrangian subspace and $\eta : \Omega \to \mathbb{F}_d$ linear) and Wigner operators ($\Omega = \mathbb{F}_d^n$ and $\eta : \Omega \to \mathbb{F}_d$ is linear). The latter have been shown to be vertices of Λ_n for any odd dimension [VFGE12, ZORH21].

Our main result will be that operators $A_\Omega^\eta \in \Lambda_n$ of the above form exhibit a very special structure. That is,

(i) Ω is a subspace and η is a linear function $\eta : \Omega \to \mathbb{F}_d^{2n}$, or

(ii) Ω is of the form

$$\Omega = \langle I, h_1 \rangle \cup \langle I, h_2 \rangle \cup \cdots \cup \langle I, h_\ell \rangle, \tag{1.6}$$

where $[h_i, h_j] \neq 0$ for all $i \neq j$ and I is an isotropic subspace with $h_i \in I^\perp$ for $i = 1, \ldots, \ell$. where $[h_i, h_j] \neq 0$ for all $i \neq j$ and I is an isotropic subspace with $h_i \in I^\perp$ for $i = 1, \ldots, \ell$. In addition, the restriction $\eta_{|\langle h_i, I \rangle}$ of η to the isotropic subspace $\langle h_i, I \rangle$ is linear for all $i = 1, \ldots, \ell$.

We conjecture that $A_\Omega^\eta \in \Lambda_n$ with Ω as in (1.6) gives a vertex of Λ_n whenever

(a) Ω is inclusion maximal among all sets of the form (1.6), i.e. there is no Ω' of the form (1.6) that strictly contains Ω,

(b) $\eta : \Omega \to \mathbb{F}_d$ cannot be extended to a linear function on $\mathbb{F}_d^{2n}$.

This conjecture is supported for two qudits with $d = 3$ by numerical computations [Zur21], and the fact that such a statement has been established for the qubit analogue of these operators [Hei19, OZR21] (see Section 4.3.1 for details).

All operators A^η_Ω of the form (1.5) contained in Λ_n should be considered as classical objects within Λ_n. This is due to the fact that each such A^η_Ω has an efficient classical description using generators of Ω. If Ω is a subspace, then A^η_Ω is fully described by a basis $a_1, \ldots, a_k$ of Ω and the images $\eta(a_1), \ldots, \eta(a_n)$. For Ω being of the form (1.6) we can fully characterize A^η_Ω by $h_1, \ldots, h_\ell$, a basis $a_1, \ldots, a_k$ of the isotropic subspace I and the images $\eta(h_i), \eta(a_j)$. As we show (Lemma 4.3.5) the maximal number of non-orthogonal elements is $dn + 1$, so again there is a classical linear description of A^η_Ω. Based on this observation, we will sketch how to update the operators A^η_Ω under Clifford unitaries and Pauli measurements efficiently classically.

The intriguing question remains open: Can we identify vertices of Λ_n or other boundary points in Λ that do not admit an efficient classical description?

Chapter 2

Stabilizer extent is not multiplicative

Stabilizer extent is not multiplicative

About this section

The following text has been previously published as:

> Arne Heimendahl, Felipe Montealegre-Mora, Frank Vallentin and David Gross. "Stabilizer extent is not multiplicative". In: *Quantum 5, 400*, 2021, `https://doi.org/10.22331/q-2021-02-24-400`

Changes from the journal version are limited to typesetting and notation. These changes were performed to match the rest of this dissertation.

Arne Heimendahl is the main contributor to this work. In particular, he developed the proofs and contributed to the presentation. David Gross sketched the overall proof idea.

Abstract

The Gottesman-Knill theorem states that a Clifford circuit acting on stabilizer states can be simulated efficiently on a classical computer. Recently, this result has been generalized to cover inputs that are close to a coherent superposition of polynomially many stabilizer states. The runtime of the classical simulation is governed by the *stabilizer extent*, which roughly measures how many stabilizer states are needed to approximate the state. An important open problem is to decide

whether the extent is multiplicative under tensor products. An affirmative answer would yield an efficient algorithm for computing the extent of product inputs, while a negative result implies the existence of more efficient classical algorithms for simulating large-scale quantum circuits. Here, we answer this question in the negative. Our result follows from very general properties of the set of stabilizer states, such as having a size that scales subexponentially in the dimension, and can thus be readily adapted to similar constructions for other resource theories.

2.1 Introduction and Summary of results

In the model of quantum computation with magic states [BK05], stabilizer circuits, whose computational power is limited by the Gottesmann-Knill theorem [Got99, SA04], are promoted to universality by implementing non-Clifford gates via the injection of *magic states*. There has been a long line of research with the goal of designing classical algorithms to simulate such circuits.

Quasiprobability-based methods [PWB15, BVDB$^+$17, HC17a, FRB18, RBVT$^+$20, HG19] work on the level of density operators. The starting point is the observation that the (qudit) Wigner function [Gro06] of stabilizer states is given by a probability distribution on phase space and thus gives rise to a classical model. Similar to the *quantum Monte-Carlo* method of many-body physics, one can then devise randomized simulation algorithms whose runtime scales with an appropriate "measure of negativity" of more general input states.

Stabilizer rank methods [BSS16, SC19, BG16], on the other hand, work with vectors in Hilbert space. The idea is to expand general input vectors as a coherent superposition of stabilizer states. The smallest number of stabilizer states required to express a given vector in this way is its *stabilizer rank*. Bravyi, Smith, and Smolin [BSS16] proposed a fast simulation algorithm. Its time complexity scales with the stabilizer rank rather than the – often much higher – dimension of the Hilbert space. Bravyi and Gosset [BG16] generalized this procedure to cover *approximate stabilizer decompositions*.

No efficient methods are known for computing the stabilizer rank analytically or numerically. To address this issue, Bravyi *et al.* [BBC$^+$19] introduced a computationally better-behaved convex relaxation: the *stabilizer extent* (see Definition 2.1.1). The central *sparsification lemma* of [BBC$^+$19] states that a stabilizer decomposition with small extent can be transformed into a sparse decomposition that is close to the original state. In this way, the stabilizer extent defines an operational measure for the degree of "non-stabilizerness". We work in a slightly more general setting than [BBC$^+$19], where the role of the stabilizer states is replaced by a finite set $\mathcal{D} \subset \mathbb{C}^d$ which spans $\mathbb{C}^d$, referred to as a *dictionary*.

Definition 2.1.1 ([BBC$^+$19])**.** *Let $\mathcal{D} \subset \mathbb{C}^d$ be a finite set of vectors spanning $\mathbb{C}^d$. For an element $x \in \mathbb{C}^d$, the* extent *of x with respect to $\mathcal{D}$ is defined as*

$$\xi_{\mathcal{D}}(x) = \min\left\{ ||c||_1^2 \,:\, c \in \mathbb{C}^{|\mathcal{D}|},\, x = \sum_{s\in\mathcal{D}} c_s s \right\},$$

where $||c||_1 = \sum_{s\in\mathcal{D}} |c_s|$. If $d = 2^n$ and $\mathcal{D} = \mathrm{STAB}_n$ is the set of stabilizer states, then $\xi_{\mathcal{D}}(x)$ is the stabilizer extent *of x, and the notation is shortened to $\xi(x)$.*

As is widely known, ℓ_1-minimizations such as $\xi_{\mathcal{D}}$ can be formulated as convex optimization problems (see for example [BV04]). In the complex case this is a *second order cone problem* [AG03], whose complexity scales polynomially in $\max(d, |\mathcal{D}|)$. In particular, the complexity of determining the stabilizer extent of an arbitrary vector, $\xi(x)$, scales exponentially in the number of qubits. Thus, the question arises whether it is possible to simplify the computation of $\xi_{\mathcal{D}}$ for certain inputs, e.g. product states of the form $\psi = \otimes_j \psi_j$.

Since the set of stabilizer states is closed under taking tensor products, one can easily see that the stabilizer extent is submultiplicative, that is $\xi(\otimes_j \psi_j) \leq \prod_j \xi(\psi_j)$ for any input state $\otimes_j \psi_j$. Bravyi *et al.* proved that it is actually multiplicative if the factors are composed of 1-, 2- or 3-qubit states.

Our main result is that stabilizer extent is *not* multiplicative in general. In fact, our result does not depend on the detailed structure

of stabilizer states, but holds for fairly general families of dictionaries. The properties used — prime among them that the size of the dictionaries scales subexponentially with the Hilbert space dimension — are listed as Properties (i) to (v) in the following theorem.

Theorem 2.1.2. *Let $(\mathcal{D}_n)$ be a sequence of dictionaries with $\mathcal{D}_n \subset (\mathbb{C}^{d_0})^{\otimes n}$ and $\mathcal{D}_1 \subset \mathbb{C}^{d_0}$ for some fixed integer d_0. Assume that $(\mathcal{D}_n)$ satisfies the following properties:*

(i) Normalization: $\langle s, s\rangle = 1$ for all $s \in \mathcal{D}_n$.

(ii) Subexponential size:

$$\log_{d_0} |\mathcal{D}_n| \leq o\left(\sqrt{d_0^n}\right).$$

(iii) Closed under complex conjugation: if $s \in \mathcal{D}_n$, then $s^* \in \mathcal{D}_n$.

(iv) Closed under taking tensor products:

$$\mathcal{D}_{n_1} \otimes \mathcal{D}_{n_2} := \{s_1 \otimes s_2 \,:\, s_1 \in \mathcal{D}_{n_1},\, s_2 \in \mathcal{D}_{n_2}\} \subset \mathcal{D}_{n_1+n_2}.$$

(v) Contains the maximally entangled state: For every n, the maximally entangled state

$$\Phi = \frac{1}{\sqrt{d_0^n}} \sum_{k \in \mathbb{Z}_{d_0}^n} e_k \otimes e_k \in \mathcal{D}_{2n}$$

is contained in the dictionary $\mathcal{D}_{2n}$. Here, $\{e_k\}$ is the standard ("computational") basis of $(\mathbb{C}^{d_0})^{\otimes n}$.

Let $\psi \in (\mathbb{C}^{d_0})^{\otimes n}$ be a unit vector. Then

$$\Pr[\xi_{\mathcal{D}_{2n}}(\psi \otimes \psi^*) < \xi_{\mathcal{D}_n}(\psi)\xi_{\mathcal{D}_n}(\psi^*)] \geq 1 - o(1).$$

In particular, for sufficiently large n, the extent with respect to the dictionary sequence $(\mathcal{D}_n)$ is strictly *submultiplicative.*

Parts of the proof of Theorem 2.1.2 follow the proof of non-multiplicativity of the *stabilizer fidelity* [BBC+19, Lemma 10]. As a crucial extra ingredient, we carefully analyze the dual second order cone formulation

of the extent and exploit *complementary slackness* to prove the fact that the optimal *dual witness* is generically unique.

Note that the main theorem also implies that other *magic monotones* recently defined in [SRP+21] (*mixed state extent, dyadic negativity,* and *generalized robustness*) fail to be multiplicative, since they all coincide with the stabilizer extent on pure states [Reg18].

The remaining part of paper is organized as follows: In Section 2.2, we outline the geometric intuition behind the argument. The rigorous proof is given in Section 2.3. As an auxiliary result, we present an optimality condition on stabilizer extent decompositions in Section 2.4.

2.2 Proof strategy

In this section, we explain the geometric intuition behind the main result. To simplify the exposition, we present a version of the argument for real vector spaces.

We recall the convex geometry underlying the problem. In the real case, the extent can be formalized as a *basis pursuit problem*:

$$\begin{aligned} \sqrt{\xi_{\mathcal{D}}(x)} \;=\; & \min \quad \sum_{s\in\mathcal{D}} |c_s| \\ & \text{s.t.} \quad c_s \in \mathbb{R} \ (s \in \mathcal{D}), \\ & \qquad \sum_{s\in\mathcal{D}} c_s s = x. \end{aligned}$$

This type of optimization can be formulated as linear program (see e.g. [BV04, Chapter 6]). Using standard techniques we can derive its dual form (see e.g. [BV04, Chapter 5]):

$$\begin{aligned} \sqrt{\xi_{\mathcal{D}}(x)} \;=\; & \max \quad x^\top y \\ & \text{s.t.} \quad y \in \mathbb{R}^d, \\ & \qquad |s^\top y| \leq 1 \text{ for all } s \in \mathcal{D}, \end{aligned}$$

where $x^\top y := \sum_{j=1}^{d} x_j y_j$ denotes the inner product on $\mathbb{R}^d$. Let

$$M_{\mathcal{D}} = \{y \in \mathbb{R}^d \,:\, |s^\top y| \leq 1 \text{ for all } s \in \mathcal{D}\}$$

be the region of feasible points for the dual program. Since $\mathcal{D}$ is finite and contains a spanning set of $\mathbb{R}^d$, the set $M_{\mathcal{D}}$ is a polytope. The dual formulation implies that for each x, there exists a *witness* y among the vertices of $M_{\mathcal{D}}$ such that $\sqrt{\xi_{\mathcal{D}}(x)} = x^\top y$. Conversely, with each vertex $y \in M_{\mathcal{D}}$, one can associate the set of primal vectors x for which y is a witness:

$$C_y = \left\{x \in \mathbb{R}^d \,:\, \sqrt{\xi_{\mathcal{D}}(x)} = x^\top y\right\} = \operatorname{cone}\left\{(-1)^k s \,:\, s \in \mathcal{D},\, k \in \{0,1\},\, (-1)^k s^\top y = 1\right\}$$

The cone over a set M, denoted by $\operatorname{cone}\{M\}$, is simply the set of all linear combinations with non-negative coefficients of a finite set of elements in M. It is easy to see that the C_y are full-dimensional convex cones that partition $\mathbb{R}^d$ as y ranges over the vertices of $M_{\mathcal{D}}$ (see Figure 2.1 for an illustration). The cones C_y are called *normal cones* and the induced partition of $\mathbb{R}^d$ is referred to as the *normal fan* of $M_{\mathcal{D}}$, see for example [Zie95]. For $x \in \mathbb{R}^d$, define the *fidelity of x with respect to $\mathcal{D}$*

$$\sqrt{F_{\mathcal{D}}(x)} := \max_{s \in \mathcal{D}} |s^\top x|$$

as the maximal overlap of x with an element in $\mathcal{D}$ (the value $\sqrt{F_{\mathcal{D}}(x)}$ can also be viewed as the ℓ_∞-norm of x with respect to $\mathcal{D}$).

These notions allow us to analyze how the extent of a vector x changes when a word w is added to the dictionary $\mathcal{D}$ (in the proof below, we will track the extent when the maximally entangled state is added to a product dictionary). Indeed, if x is contained in the interior of some C_y, and if $|w^\top y| > 1$, then the vertex y is infeasible for the dual program with respect to the dictionary $\mathcal{D} \cup \{w\}$ (i.e., $y \notin M_{\mathcal{D}\cup\{w\}}$), and therefore $\xi_{\mathcal{D}\cup\{w\}}(x) < \xi_{\mathcal{D}}(x)$.

Now, the argument of the proof of the main theorem proceeds in two steps:

(1) Assume x is chosen Haar-randomly from the unit-sphere in $\mathbb{R}^d$. Almost surely, there will be a *unique* witness y, i.e., x will lie in the interior of some normal cone C_y for some vertex y of $M_{\mathcal{D}}$. Moreover, the norm of y is large with high probability, $\|y\|_2^2 \approx$

$O(d)$. To see why the latter holds, note that

$$\|y\|_2^2 \geq (x^\top y)^2 = \xi(x) \geq \frac{1}{F_{\mathcal{D}}(x)},$$

where the second inequality follows because $x/\sqrt{F_{\mathcal{D}}(x)} \in M_D$ is feasible for the dual (as realized in [BBC+19]). A standard concentration-of-measure argument (as in [BBC+19], proof of Claim 2) shows that if $|\mathcal{D}|$ is not too large, the *maximal* inner product-squared of x with any element of $\mathcal{D}$ will be close to the expected inner product-squared with any *fixed* unit vector v, which is $|x^\top v|^2 \approx 1/d$.

(2) Now consider $x \otimes x$. With respect to the *product dictionary* $\mathcal{D} \otimes \mathcal{D}$, one easily finds that $\xi_{\mathcal{D}\otimes\mathcal{D}}(x \otimes x) = \xi_{\mathcal{D}}(x)\xi_{\mathcal{D}}(x)$, and that $y \otimes y$ is a unique witness and a vertex of $M_{\mathcal{D}\otimes\mathcal{D}}$. If Φ is the maximally entangled state,

$$\Phi^\top(y \otimes y) = d^{-1/2}\|y\|_2^2 = O(d^{1/2}) > 1.$$

Thus adding Φ to the dictionary means that $y \otimes y$ becomes dually *infeasible* (i.e., $y \otimes y \notin M_{\mathcal{D}\otimes\mathcal{D}\cup\{\Phi\}}$). It follows that the extent of $x \otimes x$ (in fact, the extent of *any* element in the interior of $C_{y\otimes y}$) decreases if Φ is added.

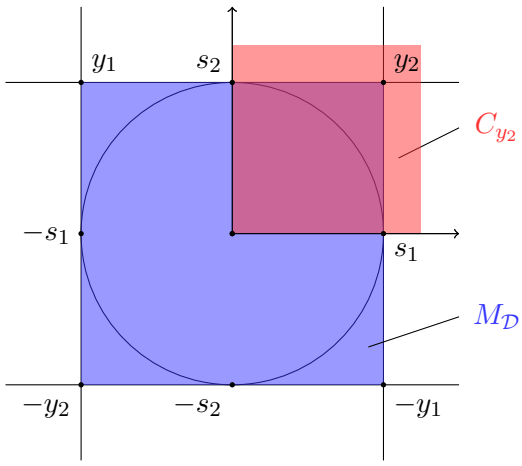


Figure 2.1: The polytope $M_{\mathcal{D}}$ for the dictionary $\mathcal{D} = \{s_1, s_2\} \subset \mathbb{S}^1$ and the normal cone C_{y_2} of the vertex y_2. The active inequalities at y_2 yield the extreme rays of C_{y_2}.

2.3 Proof of the main theorem

In preparation of proving the main theorem, we translate the convex geometry of ℓ_1-minimization from the real case (sketched in the previous section) to the case of complex vector spaces. This problem has been treated before in various places in the literature, including in [BSS16], in the context of the theory of *compressed sensing* (e.g. [FR13]), and in greater generality in the convex optimization literature (e.g. [Pat00]). As we are not aware of a reference that gives a concise account of all the statements required, we present self-contained proofs in Appendix 2.A.

We will use the superscripts R and I to denote, respectively, the real and complex part of a vector. The extent then has the following dual formulation (c.f. Appendix 2.A):

$$\begin{aligned} \sqrt{\xi_{\mathcal{D}}(\psi)} \;=\; & \max \quad (\psi^R)^\top y^R + (\psi^I)^\top y^I \\ & \text{s.t.} \quad y \in \mathbb{C}^d, \\ & \qquad \sqrt{F_{\mathcal{D}}(y)} \leq 1, \end{aligned}$$

where

$$F_{\mathcal{D}}(y) = \max_{s \in \mathcal{D}} |\langle s, y \rangle|^2$$

and $\langle s, y \rangle := \sum_{j=1}^{d} \overline{s_j} y_j$ denotes the inner product on $\mathbb{C}^d$.

Let

$$M_{\mathcal{D}} = \{y \in \mathbb{C}^d \,:\, |\langle s, y \rangle| \leq 1 \text{ for all } s \in \mathcal{D}\}$$

be the set of feasible points for the dual. In contrast to the real case, $M_{\mathcal{D}}$ is not a polytope, but $M_{\mathcal{D}}$ is still a bounded convex set (viewed as a subset in $\mathbb{R}^{2d}$ – for a more detailed explanation, see Appendix 2.A). Thus, by Krein-Millman, $M_{\mathcal{D}}$ is the convex hull of its extreme points, which can be characterized as follows (Appendix 2.A contains a proof):

Proposition 2.3.1. *A point $y \in M_{\mathcal{D}}$ is an extreme point of $M_{\mathcal{D}}$ if and only if*

$$\big\{s \in \mathcal{D} \,:\, |\langle s, y \rangle| = 1\big\}$$

is a spanning set for $\mathbb{C}^d$.

We will continue with an example of one extreme point of $M_{\mathcal{D}}$ for $\mathcal{D} = \mathrm{STAB}_n$ being the dictionary of n-qubit stabilizer states.

Example 2.3.2. *One extreme point for the set* $\mathcal{M}_{\mathrm{STAB}_n}$ *is the rescaled tensor-power* $\psi_T^{\otimes n}/F(\psi_T^{\otimes n})$ *of the magic* T*-state,*

$$\psi_T := \begin{pmatrix} \cos(\beta) \\ e^{i\frac{\pi}{4}} \sin(\beta) \end{pmatrix},$$

where $\beta = \frac{1}{2}\arccos(\frac{1}{\sqrt{3}})$. *In this remark, we sketch why this is so.*

The vector $\psi_T^{\otimes n}$ *satisfies* $\xi(\psi_T^{\otimes n}) = 1/F(\psi_T^{\otimes n})$ *[BBC+19, Proposition 2]. Now,* $\psi_T\psi_T^\dagger = \frac{1}{3}(\mathbb{I} + C + C^2)$ *where* C *is the Clifford matrix which cyclically permutes the Pauli matrices* $\{X, Y, Z\}$. *This way, if* $U = C^{i_1} \otimes \cdots \otimes C^{i_n}$, *then*

$$\langle U^\dagger s, \psi_T^{\otimes n}\rangle = \langle s, U\psi_T^{\otimes n}\rangle = \langle s, \psi_T^{\otimes n}\rangle \qquad \text{for all } s \in \mathrm{STAB}_n.$$

It follows that the group generated by tensor products of $\{\mathbb{I}, C\}$ *acts on the optimizers of* $F(\psi_T^{\otimes n})$. *But the standard basis vector* $e_0^{\otimes n}$ *is one such optimizer [BBC+19, Lemma 2] and*

$$\mathrm{Span}\{e_0,\, Ce_0\} = \mathrm{Span}\{e_0,\, (e_0 + e_1)/\sqrt{2}\} = \mathbb{C}^2.$$

This shows that the optimizers of $F(\psi_T^{\otimes n})$ *contain all tensor products of* e_0 *and* $(e_0 + e_1)/\sqrt{2}$, *which form a basis for* $(\mathbb{C}^2)^{\otimes n}$.

Finally, $\psi_T^{\otimes n}/F(\psi_T^{\otimes n})$ *is an optimal dual witness for* $\psi_T^{\otimes n}$. *By Prop. 2.3.1, then, this witness is extremal.*

Returning back to the general theory, we associate a *normal cone* with every extreme point y:

$$\begin{aligned} C_y &= \left\{\psi \in \mathbb{C}^d \,:\, \langle \psi, y\rangle^R = \max_{p \in M_{\mathcal{D}}} \langle \psi, p\rangle^R\right\} \\ &= \mathrm{cone}\left\{e^{i\phi}s \,:\, s \in \mathcal{D}, \phi \in \mathbb{R},\, e^{i\phi}\langle s, y\rangle = 1\right\}. \end{aligned} \tag{2.1}$$

Notice that

$$\langle \psi, y\rangle^R = (\psi^R)^\top y^R + (\psi^I)^\top y^I.$$

A final preparation step invokes *complementary slackness* (Appendix 2.A contains a proof):

Lemma 2.3.3 (Complementary slackness conditions)**.** *Let $y \in M_{\mathcal{D}}$ be* any *optimal dual witness, i.e., $\psi \in C_y$ and $\sqrt{\xi_{\mathcal{D}}(\psi)} = \langle \psi, y \rangle^R$. Then for* any *optimal extent decomposition $\psi = \sum_{s\in\mathcal{D}} c_s s$ with $\sqrt{\xi_{\mathcal{D}}(\psi)} = \sum_{s\in\mathcal{D}} |c_s|$ we have the following two conditions:*

(I) If $c_s \neq 0$, then $\langle s, y\rangle = c_s/|c_s|$.

(II) If $|\langle s, y\rangle| < 1$, then $c_s = 0$.

The complementary slackness conditions have the following two consequences:

First, assume that $\psi = \sum_{s\in\mathcal{D}} c_s s$ is an optimal decomposition and that $y \in \mathbb{C}^d$ optimal for the dual. From condition (I), we obtain

$$|\langle \psi, y\rangle| = \left|\sum_{s\in\mathcal{D}} \overline{c_s}\langle s, y\rangle\right| = \left|\sum_{s\in\mathcal{D}, c_s\neq 0} \overline{c_s}\frac{c_s}{|c_s|}\right| = \sum_{s\in\mathcal{D}} |c_s| = \sqrt{\xi_{\mathcal{D}}(\psi)},$$

so we can rewrite the dual program for the extent as

$$\begin{aligned} \xi_{\mathcal{D}}(\psi) \;=\; & \max \quad |\langle \psi, y\rangle|^2 \\ & \text{s.t.} \quad y \in \mathbb{C}^d, \\ & \qquad F_{\mathcal{D}}(y) \leq 1, \end{aligned} \tag{2.2}$$

which coincides with the dual formulation given in [BBC$^+$19]. Since $\psi/\sqrt{F_{\mathcal{D}}(\psi)}$ is feasible for the dual, we get the natural lower bound [BBC$^+$19]

$$\xi_{\mathcal{D}}(\psi) \geq \frac{1}{F_{\mathcal{D}}(\psi)}. \tag{2.3}$$

Secondly, if a state ψ is chosen Haar-randomly, the optimal dual witness y for $\xi_{\mathcal{D}}(\psi)$ is an extreme point *and* unique of $M_{\mathcal{D}}$ with probability one, because of the following observation: A generic ψ will not be contained in a proper subspace spanned by elements of $\mathcal{D}$, since the finite collection of all these lower-dimensional subspaces has measure zero. Thus, generically, if we expand $\psi = \sum_{s\in\mathcal{D}} c_s s$ in the dictionary $\mathcal{D}$, the set $\{s \in \mathcal{D} : c_s \neq 0\}$ has to span $\mathbb{C}^d$.

Now suppose we are given two optimal dual witnesses y_1, y_2 for $\xi_{\mathcal{D}}(\psi)$. Condition (I) of 2.3.3 tells us that for *all* optimal primal extent

decompositions, both y_1 and y_2 are solutions of the system of linear equations:

$$\langle s, y\rangle = \frac{c_s}{|c_s|} \quad \text{for all } c_s \neq 0.$$

However, this system has a unique solution because the words $s \in \mathcal{D}$ with $c_s \neq 0$ span $\mathbb{C}^d$ and therefore, $y_1 = y_2$. Such ψ's are also called *non-degenerate* in convex optimization [AG03].

Analogously to the case of a normal cone in a real-valued vector space, note that the interior $\mathrm{int}(C_y)$ of a normal cone C_y consists of all points ψ whose dual witness is unique and the extreme point y. This means that there exists an optimal extent decomposition

$$\psi = \sum_{s\in\mathcal{D}} c_s s = \sum_{s\in\mathcal{D}} \alpha_s\, e^{i\phi_s} s,$$

such that

$$\alpha_s \geq 0,\ c_s = \alpha_s e^{i\phi s},\ e^{i\phi_s} s \in C_y, \text{ and } \{s \in \mathcal{D} : c_s \neq 0\} \text{ spans } \mathbb{C}^d.$$

With the above notion, we are able to describe how the extent is effected by adding a word w to the dictionary $\mathcal{D}$. As in the case of a real valued vector space, an extreme point $y \in M_{\mathcal{D}}$ becomes dually infeasible if $|\langle w, y\rangle| > 1$ (i.e., $y \notin M_{\mathcal{D}\cup\{w\}}$). Hence, the extent of an element x decreases if y is the unique dual witness of x, that is $x \in \mathrm{int}(C_y)$. In summary, we get the following theorem:

Theorem 2.3.4. *Let $\mathcal{D} \subset \mathbb{C}^d$ be a dictionary and let $w \in \mathbb{C}^d$ with $\langle w, w\rangle = 1$. Let $\mathcal{D}' = \mathcal{D} \cup \{w\}$. Then, $\xi_{\mathcal{D}'}(x) < \xi_{\mathcal{D}}(x)$, if and only if $x \in \mathrm{int}(C_y)$ for an extreme point $y \in M_{\mathcal{D}}$ with $|\langle w, y\rangle| > 1$.*

In order to analyze the multiplicativity properties of the extent for product inputs, we now turn our attention to *product dictionaries*. The argument starts with the observation that extreme points of $M_{\mathcal{D}}$ are closed under taking tensor products. That is, if y_1, y_2 are extreme points of dually feasible sets $M_{\mathcal{D}_j} \subset \mathbb{C}^{d_j}$ for two dictionaries $\mathcal{D}_1$ and $\mathcal{D}_2$, then $y_1 \otimes y_2$ is an extreme point of $M_{\mathcal{D}_1\otimes\mathcal{D}_2}$, where $\mathcal{D}_1 \otimes \mathcal{D}_2 \subset \mathbb{C}^{d_1} \otimes \mathbb{C}^{d_2}$ is the product dictionary. Indeed, since $y_1 \otimes y_2 \in M_{\mathcal{D}_1\otimes\mathcal{D}_2}$ and the set

$$\begin{aligned} &\{s_1 \otimes s_2 \in \mathcal{D}_1 \otimes \mathcal{D}_2 \,:\, |\langle s_1 \otimes s_2, y_1 \otimes y_2\rangle| = 1\} \\ = \ &\{s_1 \otimes s_2 \in \mathcal{D}_1 \otimes \mathcal{D}_2 \,:\, |\langle s_j, y_j\rangle| = 1,\ j = 1, 2\} \end{aligned}$$

is a spanning set of $\mathbb{C}^{d_1} \otimes \mathbb{C}^{d_2}$. Moreover, by the characterization of the normal cone (2.1), it follows immediately that the normal cone of $y_1 \otimes y_2$ has the form

$$C_{y_1 \otimes y_2} = \operatorname{cone}\{e^{i\phi_{s_1}} s_1 \otimes e^{i\phi_{s_2}} s_2 \,:\, e^{i\phi_{s_j}} s_j \in C_{y_j},\, j = 1, 2\}. \tag{2.4}$$

This allows us to derive the following multiplicativity property of product dictionaries:

Lemma 2.3.5. *Consider two dictionaries $\mathcal{D}_j \subset \mathbb{C}^{d_j}$ and extreme points $y_j \in M_{\mathcal{D}_j}$, $j = 1, 2$. Then, $C_{y_1} \otimes C_{y_2} \subset C_{y_1 \otimes y_2}$ and $\operatorname{int}(C_{y_1}) \otimes \operatorname{int}(C_{y_2}) \subset \operatorname{int}(C_{y_1 \otimes y_2})$. Therefore,*

$$\xi_{\mathcal{D}_1 \otimes \mathcal{D}_2}(\psi_1 \otimes \psi_2) = \xi_{\mathcal{D}_1}(\psi_1) \xi_{\mathcal{D}_2}(\psi_2)$$

for all $\psi_j \in \mathbb{C}^{d_i}$.

Proof. We will prove $C_{y_1} \otimes C_{y_2} \subset C_{y_1 \otimes y_2}$, the statement $\operatorname{int}(C_{y_1}) \otimes \operatorname{int}(C_{y_2}) \subset \operatorname{int}(C_{y_1 \otimes y_2})$ can be proven analogously. Let $\psi_j \in C_j$, so

$$\psi_j = \sum_{s \in \mathcal{D}} \alpha_s^j \, e^{i\phi_s^j} s,$$

where $\alpha_s^j \geq 0$ and if α_s^j is positive, then $e^{i\phi_s^j} s \in C_{y_j}$. Thus,

$$\psi_1 \otimes \psi_2 = \sum_{s \otimes s' \mathcal{D}_1 \otimes \mathcal{D}_2} \alpha_s^1 \alpha_{s'}^2 \, (e^{i\phi_s^1} s \otimes e^{i\phi_{s'}^2} s') \in C_{y_1 \otimes y_2},$$

by Equation (2.4).

In order to prove multiplicativity it suffices to observe that, by the definition of the normal cone and the extent formulation (2.2),

$$\xi_{\mathcal{D}_1 \otimes \mathcal{D}_2}(\psi_1 \otimes \psi_2) = |\langle \psi_1 \otimes \psi_2, y_1 \otimes y_2 \rangle|^2 = |\langle \psi_1, y_1 \rangle|^2 \, |\langle \psi_2, y_2 \rangle|^2 = \xi_{\mathcal{D}_1}(\psi_1) \xi_{\mathcal{D}_2}(\psi_2).$$

□

Using the above lemma and the generic uniqueness of the dual witness y, we are now able to prove our main theorem. We subdivide the proof in two parts, where the first part is an adaption of Claim 2 in [BBC$^+$19] to the class of dictionaries defined in Theorem 2.1.2:

Proposition 2.3.6. *Assume that the dictionary sequence $(\mathcal{D}_n)$ with $\mathcal{D}_n \subset (\mathbb{C}^{d_0})^{\otimes n}$ satisfies the assumptions of Theorem 2.1.2. Then, for a Haar-randomly chosen unit vector $\psi \in (\mathbb{C}^{d_0})^{\otimes n}$ and some fixed $\varepsilon > 0$ it holds that*

$$\Pr\left[F_{\mathcal{D}_n}(\psi) \leq \frac{1}{\sqrt{d_0^n} + \varepsilon}\right] \geq 1 - o(1).$$

In particular, $F_{\mathcal{D}_n}(\psi) \leq \frac{1}{\sqrt{d_0^n}+\varepsilon}$ for sufficiently large n and a typical unit vector $\psi \in (\mathbb{C}^{d_0})^{\otimes n}$.

Proof. We fix a unit vector $\omega \in (\mathbb{C}^{d_0})^{\otimes n}$ and choose a Haar-random unit vector $\psi \in (\mathbb{C}^{d_0})^{\otimes n}$. Following the proof of Claim 2 in [BBC+19] we can bound the probability of the event $\{|\langle \omega, \psi\rangle|^2 \geq x\}$ by

$$\Pr[|\langle \omega, \psi\rangle|^2 \geq x] = (1-x)^{d_0^n - 1} \leq e^{-x(d_0^n - 1)}.$$

If we set $x = (\sqrt{d_0^n} + \varepsilon)^{-1}$ for $\varepsilon > 0$ and use Properties (i) and (ii), we can use a union bound to estimate the fidelity of ψ with respect to $\mathcal{D}_n$ by

$$\begin{aligned}\Pr\left[\max_{s\in\mathcal{D}} |\langle \psi, s\rangle|^2 \geq \frac{1}{\sqrt{d_0^n} + \varepsilon}\right] &\leq |\mathcal{D}_n| \cdot \exp\left(-\frac{d_0^n - 1}{\sqrt{d_0^n} + \varepsilon}\right) \\ &\leq \exp\left(o\left(\sqrt{d_0^n}\right)\ln(d_0) - \frac{d_0^n - 1}{\sqrt{d_0^n} + \varepsilon}\right),\end{aligned}$$

which converges to zero as n tends to infinity. □

The proposition assures that randomly chosen unit vectors generically have small overlap with elements in the dictionary sequence. Starting from there, we proceed with the proof of the main theorem.

Proof of Theorem 2.1.2. Let $\psi \in (\mathbb{C}^{d_0})^{\otimes n}$ be a unit vector satisfying $F_{\mathcal{D}_n}(\psi) \leq \frac{1}{\sqrt{d_0^n}+\varepsilon}$ for some $\varepsilon > 0$. Due to Proposition 2.3.6, this holds for a typical ψ and sufficiently large n. As a consequence of (2.3), we can lower bound the extent of ψ by

$$\xi_{\mathcal{D}_n}(\psi) \geq \frac{1}{F_{\mathcal{D}_n}(\psi)} \geq \sqrt{d_0^n} + \varepsilon.$$

Let $y \in M_{\mathcal{D}_n}$ be an optimal dual witness, so $\psi \in C_y$. As pointed out earlier, we can further assume that y is an extreme point of $M_{\mathcal{D}_n}$ and that $y \in \operatorname{int}(C_y)$ generically. Applying Cauchy-Schwarz, we get a lower bound on the norm of y by

$$|\langle y, y\rangle| = |\langle y, y\rangle| \cdot |\langle \psi, \psi\rangle| \geq |\langle \psi, y\rangle|^2 = \xi_{\mathcal{D}_n}(\psi) \geq \sqrt{d_0^n} + \varepsilon. \tag{2.5}$$

Now consider $\psi \otimes \psi^*$. Assumption (iii) ensures that $\xi_{\mathcal{D}}(\psi) = \xi_{\mathcal{D}}(\psi^*)$ and $\psi^* \in \operatorname{int}(C_{y^*})$. The proof of Lemma 2.3.5 tells us that the extreme point $y \otimes y^*$ of $M_{\mathcal{D}_n \otimes \mathcal{D}_n}$ is optimal for

$$\xi_{\mathcal{D}_n \otimes \mathcal{D}_n}(\psi \otimes \psi^*) = \xi_{\mathcal{D}_n}(\psi)\xi_{\mathcal{D}_n}(\psi^*).$$

Moreover, it is the unique optimizer, as $\psi \otimes \psi^* \in \operatorname{int}(C_y) \otimes \operatorname{int}(C_{y^*}) \subset \operatorname{int}(C_{y \otimes y^*})$.

Next, we add the maximally entangled state Φ to the dictionary and observe

$$\xi_{\mathcal{D}_n \otimes \mathcal{D}_n}(\psi \otimes \psi^*) \geq \xi_{\mathcal{D}_n \otimes \mathcal{D}_n \cup \{\Phi\}}(\psi \otimes \psi^*),$$

since $\mathcal{D}_n \otimes \mathcal{D}_n \subset \mathcal{D}_n \otimes \mathcal{D}_n \cup \{\Phi\}$. The norm estimation (2.5) of y yields

$$\begin{aligned}\max_{s \in \mathcal{D}_n \otimes \mathcal{D}_n \cup \{\Phi\}} |\langle s, y \otimes y^*\rangle|^2 \geq |\langle \Phi, y \otimes y^*\rangle|^2 &= \Big|\frac{1}{\sqrt{d}} \sum_{k \in \mathbb{Z}_{d_0}^n} \langle y, e_k\rangle\langle y^*, e_k\rangle\Big|^2 \\ &= \frac{1}{d}|\langle y, y\rangle|^2 > 1,\end{aligned}$$

therefore $y \otimes y^*$ is not contained in the set of dually feasible points $M_{\mathcal{D}_n \otimes \mathcal{D}_n \cup \{\phi\}}$ of the dictionary $\mathcal{D}_n \otimes \mathcal{D}_n \cup \{\phi\}$. Since $y \otimes y^* \in \operatorname{int}(C_{y \otimes y^*})$ we can apply Theorem 2.3.4 to obtain $\xi_{\mathcal{D}_n \otimes \mathcal{D}_n \cup \{\Phi\}}(\psi \otimes \psi^*) < \xi_{\mathcal{D}_n \otimes \mathcal{D}_n}(\psi \otimes \psi^*)$.

To conclude, because of (iv) and (v),

$$\xi_{\mathcal{D}_{2n}}(\psi \otimes \psi^*) \leq \xi_{\mathcal{D}_n \otimes \mathcal{D}_n \cup \{\Phi\}}(\psi \otimes \psi^*) < \xi_{\mathcal{D}_n \otimes \mathcal{D}_n}(\psi \otimes \psi^*) = \xi_{\mathcal{D}_n}(\psi)\xi_{\mathcal{D}_n}(\psi^*),$$

which proves the desired result. □

2.4 An optimality condition for the stabilizer extent

In this section we fix the dictionary sequence to be the set of n-qubit stabilizer states STAB_n and we will derive a condition on optimal

stabilizer extent decompositions. (While preparing this document, we learned that this fact had already been observed earlier [Cam20], but it does not seem to be published).

Let $P_n = \left\{ \otimes_{i=1}^n W_i \,:\, W_i \in \{I, X, Y, Z\} \right\}$ be the set of n-qubit Pauli matrices. The set of stabilizer states can be decomposed in a disjoint union of orthonormal bases, where each basis is labeled by a maximally commuting set $\mathcal{S} \subset P_n$ of Pauli matrices (see [NC11], Chapter 10, or [Gro06, KG15] for details). The projectors on the basis elements can be written as $ss^\dagger = \frac{1}{2^n} \sum_{\sigma \in \mathcal{S}} (-1)^{k_\sigma} \sigma$, where $k_\sigma \in \{0, 1\}$ has to be chosen in a way such that $\{(-1)^{k_\sigma} \sigma \,:\, \sigma \in \mathcal{S}\}$ is a closed matrix group with 2^n elements.

Theorem 2.4.1. *Let ψ be an n-qubit state. Suppose that $\psi = \sum c_s s$ is an optimal stabilizer extent decomposition, that is $\xi(\psi) = \left(\sum_{s \in \mathcal{D}} |c_s| \right)^2$. Then there is at most one non-zero c_s for the words s that are labeled by the same orthonormal basis.*

For the proof of the theorem, we will make use of the *Clifford group* $\mathcal{C}_n$. For our purpose this is the unitary group that preserves the set STAB_n, i.e., if $U \in \mathcal{C}_n$, then $Us \in \mathrm{STAB}_n$ for all $s \in \mathrm{STAB}_n$ (more details can be found in [Gro06]).

Proof. First, we prove the statement for the 1-qubit case. The 1-qubit stabilizer dictionary is given by the disjoint union of three orthonormal bases

$$\mathrm{STAB}_1 = \mathcal{B}_1 \,\dot\cup\, \mathcal{B}_2 \,\dot\cup\, \mathcal{B}_3,$$

where the three orthonormal stabilizer bases are given by

$$\mathcal{B}_1 = \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\}, \quad \mathcal{B}_2 = \left\{ \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ i \end{pmatrix}, \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -i \end{pmatrix} \right\}, \quad \mathcal{B}_3 = \left\{ \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right\}.$$

Because the Clifford group acts transitively on $\{\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3\}$ and maps optimal decompositions to optimal decompositions – i.e. if $\psi = \sum_{s \in \mathrm{STAB}_n} c_s s$ is optimal, then so is $U\psi = \sum_{s \in \mathrm{STAB}_n} c_s (Us)$ – it suffices to prove the statement for a single basis, e.g. $\mathcal{B}_1$.

So suppose that we have decomposition of some state $\psi = \sum_{s \in \mathrm{STAB}_1} c_s s$ with non-negative coefficients in the basis $\mathcal{B}_1$. Since optimal ℓ_1-decompositions are invariant under scaling with a complex number, we may assume that the part of the decomposition realized by $\mathcal{B}_1$ is of the form

$$\omega = 1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + z \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \text{or} \quad \omega = z \begin{pmatrix} 1 \\ 0 \end{pmatrix} + 1 \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

with $z = x + iy \in \mathbb{C}$ and $|x| + |y| \leq 1$. Hence, the coefficients have ℓ_1-norm $1 + |z| = 1 + \sqrt{x^2 + y^2}$. If ω is of the first form, then we can also decompose it as

$$\omega = (\sqrt{2}|x|) \cdot \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ \mathrm{sign}(x) \end{pmatrix} + (\sqrt{2}|y|) \cdot \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ \mathrm{sign}(y)i \end{pmatrix} + (1 - |x| - |y|) \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{2.6}$$

and the ℓ_1-norm of the coefficients in this decomposition is

$$\sqrt{2}|x| + \sqrt{2}|y| + (1 - |x| - |y|) = 1 + (\sqrt{2} - 1)(|x| + |y|) < 1 + \frac{1}{2}(|x| + |y|).$$

But

$$\sqrt{(\xi(\omega))} \leq 1 + \frac{1}{2}(|x| + |y|) < 1 + \sqrt{x^2 + y^2} = 1 + |z|,$$

so the decomposition of ω using only elements of $\mathcal{B}_1$ is not optimal. By changing the two coordinates, we can argue analogously if $\omega = \begin{pmatrix} z \\ 1 \end{pmatrix}$. Updating the decomposition of ψ to $\psi = \sum_{s \in \mathrm{STAB}_1} \hat{c}_s s$ via the new decomposition of ω (2.6), we also get a new decomposition of ψ with lower ℓ_1-norm and only one non-zero coefficient for the basis $\mathcal{B}_1$. This follows by comparing $\sum_{s \in \mathrm{STAB}_1} |c_s|$ with $\sum_{s \in \mathrm{STAB}_1} |\hat{c}_s|$ via the triangle inequality.

For the n-qubit case assume that $\psi = \sum c_s s$ is a stabilizer decomposition with $c_s c_{s'} \neq 0$ for two stabilizer states $s, s' \in \mathrm{STAB}_n$ belonging to the same orthonormal basis. Due to invariance of ξ under the Clifford group and its transitive action on orthonormal stabilizer bases, we may choose any orthonormal stabilizer basis. By possibly applying another Clifford unitary, we may even assume that

$s = e_0 \otimes e_0 \cdots e_0$, $s' = e_1 \otimes e_0 \cdots e_0$. But if we consider the decomposition of the unnormalized state

$$\omega = c_s e_0 \otimes e_0 \otimes \cdots \otimes e_0 + c_{s'} e_1 \otimes e_0 \otimes \cdots \otimes e_0 = (c_s e_0 + c_{s'} e_1) \otimes e_0 \otimes \cdots \otimes e_0,$$

the 1-qubit case result together with the fact that stabilizer states are closed under taking tensor products can be applied to see that the decomposition of ω is not optimal. Now, the crucial observation is that if $\psi = \sum c_s s$ is an optimal stabilizer extent decomposition, then $\omega = c_s s + c_{s'} s'$ is an optimal decomposition for ω. But as the decomposition of ω is not optimal, neither is the one of ψ. □

There is an interesting connection between the derived optimality condition and the geometric properties of the *stabilizer polytope* SP_n, which is the convex hull of the projectors onto stabilizer states, i.e., $SP_n = \mathrm{conv}\{ss^\dagger : s \in \mathrm{STAB}_n\}$. As shown in [Hei19, EG15], two stabilizer projectors are connected by an edge if and only if they do not belong to the same orthonormal stabilizer basis. Thus, we can reformulate the above result:
If $\psi = \sum c_s s$ is an optimal stabilizer extent decomposition and $c_s c_{s'} \neq 0$, then the set $\mathrm{conv}\{ss^\dagger, s'(s')^\dagger\}$ is an edge of SP_n.

Summary and outlook

We have settled an open problem in stabilizer resource theory, by showing that the stabilizer extent is generically sub-multiplicative in high dimensions. What is striking is that the previous multiplicativity results for one to three qubit states [BSS16] made use of the detailed structure of the set of stabilizer states. In contrast, our counterexample involves only a small number of high-level properties of the stabilizer dictionary. Therefore, we see this work as evidence that ℓ_1-based complexity measures on tensor product spaces should be expected to be strictly sub-multiplicative in the absence of compelling reasons to believe otherwise. In particular, it seems highly plausible that the assumptions that go into Theorem 2.1.2 can be considerably weakened. We leave this problem open for future analysis.

Acknowledgments

We thank Markus Heinrich, Earl Campbell, Richard Küng, and James Seddon for interesting discussions and feedback.

This work has been supported by the DFG (SPP1798 CoSIP), Germany's Excellence Strategy – Cluster of Excellence *Matter and Light for Quantum Computing* (ML4Q) EXC2004/1, Cologne's Key Profile Area *Quantum Matter and Materials*, and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie agreement No 764759. The third named author is partially supported by the SFB/TRR 191 "Symplectic Structures in Geometry, Algebra and Dynamics" and by the project "Spectral bounds in extremal discrete geometry" (project number 414898050), both funded by the DFG.

2.A Formulating the extent as a second order cone program

Here, we write the extent of Definition 2.1.1 with respect to a complex dictionary $\mathcal{D} \subset \mathbb{C}^d$ as a real second order cone program in standard form [AG03]. We impose the condition that the elements in $\mathcal{D}$ are normalized, i.e., $\langle w, w\rangle = 1$. For an optimal decomposition $\psi = \sum_{s\in\mathcal{D}} c_s s$ we set $c_s^R = \operatorname{Re} c_s$ and $c_s^I = \operatorname{Im} c_s$. The standard primal version of the extent is given by

$$\begin{aligned}\sqrt{\xi_{\mathcal{D}}(\psi)} \;=\; \min \quad & \sum_{s\in\mathcal{D}} t_s \\ \text{s.t.} \quad & \sum_{s\in\mathcal{D}} \begin{bmatrix} s^R & -s^I & 0 \\ s^I & s^R & 0 \end{bmatrix} \cdot \begin{bmatrix} c_s^R \\ c_s^I \\ t_s \end{bmatrix} = \begin{bmatrix} \psi^R \\ \psi^I \end{bmatrix} \\ & (c_s^R, c_s^I, t_s) \in \mathcal{L}^{2+1}\,(s \in \mathcal{D}),\end{aligned}$$

where

$$\mathcal{L}^{2+1} = \left\{(x_1, x_2, t) \in \mathbb{R}^3 \,:\, \sqrt{x_1^2 + x_2^2} \leq t\right\}$$

is the 3-dimensional Lorentz cone. As the program is in primal standard form, we can derive its dual formulation:

$$\begin{aligned} \max \quad & (\psi^R)^\top y^R + (\psi^I)^\top y^I \\ \text{s.t.} \quad & \begin{bmatrix} (s^R)^\top & (s^I)^\top \\ (-s^I)^\top & (s^R)^\top \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} y^R \\ y^I \end{bmatrix} + z_s = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \text{ for all } s \in \mathcal{D}, \\ & z_s \in \mathcal{L}^{2+1}\,(s \in \mathcal{D}),\ (y^R, y^I) \in \mathbb{R}^{2d}. \end{aligned} \tag{2.7}$$

Since $\mathcal{D}$ contains a basis of $\mathbb{C}^d$, both programs are strictly feasible and strong duality holds, so the optimal values for min and max coincide. The dual constraints are equivalent to $\max_{s\in\mathcal{D}} |\langle s, y\rangle| \leq 1$, where $y = y^R + iy^I \in \mathbb{C}^d$. Thus, we can rewrite the dual as

$$\begin{aligned} \max \quad & (\psi^R)^\top y^R + (\psi^I)^\top y^I \\ \text{s.t.} \quad & |\langle s, y\rangle| \leq 1 \text{ for all } s \in \mathcal{D}, \\ & y \in \mathbb{C}^d. \end{aligned}$$

Next, we prove Proposition 2.3.1, which gives a characterization of the extreme points of the set of dually feasible points $M_\mathcal{D} = \{y \in \mathbb{C}^d : |\langle s, y\rangle| \leq 1 \text{ for all } s \in \mathcal{D}\}$.

Proof of Proposition 2.3.1. Let $y \in M_\mathcal{D}$. First, we assume that the set

$$A_y = \{s \in \mathcal{D} : |\langle s, y\rangle| = 1\}$$

does not span $\mathbb{C}^d$. Then, there exists $u \in \mathbb{C}^d$ being orthogonal to all elements in A_y and, since d is finite, we can find $\varepsilon > 0$ such that $y \pm \varepsilon u \in M_\mathcal{D}$ and $y = \frac{1}{2}((y + \varepsilon u) + (y - \varepsilon u))$ is a proper convex combination of $y \pm \varepsilon$. Hence, y is not an extreme point of $M_\mathcal{D}$.

Conversely, assume that A_y spans $\mathbb{C}^d$ and that $y = \alpha u + (1-\alpha)v$ for some $u, v \in M_\mathcal{D}$. For every $s \in A_y$ there is $\phi_s \in \mathbb{R}$ such that

$$1 = e^{i\phi_s}\langle s, y\rangle = \alpha e^{i\phi_s}\langle s, u\rangle + (1-\alpha)e^{i\phi_s}\langle s, v\rangle,$$

hence,

$$\left(e^{i\phi_s}\langle s, u\rangle\right)^R = \left(e^{i\phi_s}\langle s, v\rangle\right)^R = 1.$$

But as $|\langle s,u\rangle| \leq 1$ and $|\langle s,v\rangle| \leq 1$, it must hold that

$$\left(e^{i\phi_s}\langle s,u\rangle\right)^I = \left(e^{i\phi_s}\langle s,v\rangle\right)^I = 0.$$

Since the elements of A_y span $\mathbb{C}^d$, the system

$$e^{i\phi_s}\langle s,w\rangle = 1 \text{ for all } s \in A_y \text{ and } w \in \mathbb{C}^d$$

has the unique solution y, so $y = u = v$ and y is an extreme point of M_y. □

We will continue with the proof of Lemma 2.3.3, which is a consequence of complementary slackness.

Proof of Lemma 2.3.3. If $(c_s,t_s)_{s\in\mathcal{D}}$ is optimal for the primal and $(y,(z_s)_{s\in\mathcal{D}})$ optimal for the dual, then complementary slackness [AG03] enforces

$$\sum_{s\in\mathcal{D}} (c_s,t_s)\cdot z_s = 0,$$

but as we have

$$z_s = (-\langle s,y\rangle^R, -\langle s,y\rangle^I, 1),$$

due to the duality constraint (2.7), we can rewrite this as

$$\sum_{s\in\mathcal{D}} t_s = \sum_{s\in\mathcal{D}} c_s^R\langle s,y\rangle^R + c_s^I\langle s,y\rangle^I.$$

Applying Cauchy-Schwarz to each term of the right hand side we obtain

$$\begin{aligned}\sum_{s\in\mathcal{D}} c_s^R\langle s,y\rangle^R + c_s^I\langle s,y\rangle^I &\leq \sum_{s\in\mathcal{D}} ||(c_s^R,c_s^I)||_2 \cdot ||\left(\langle s,y\rangle^R,\langle s,y\rangle^I\right)||_2 \\ &= \sum_{s\in\mathcal{D}} ||c_s||_2\cdot|\langle s,y\rangle| \\ &\leq \sum_{s\in\mathcal{D}} t_s,\end{aligned}$$

where the last inequality follows from $(c_s,t_s)\in\mathcal{L}^{2+1}$ and $|\langle s,y\rangle|\leq 1$ for all $s\in\mathcal{D}$. Consequently, we have equality in each step. This leads to the conditions given in the lemma because:

(I) If $c_s \neq 0$, then $|\langle s,y\rangle| = 1$, but by the first inequality the vector (c_s^R,c_s^I) must be proportional to $(\langle s,y\rangle^R,\langle s,y\rangle^I)$, hence $\langle s,y\rangle = \frac{c_s}{|c_s|}$.

(II) If $|\langle s,y\rangle| < 1$, then $c_s = 0$. □

Chapter 3

The axiomatic and the operational approaches to resource theories of magic do not coincide

About this section

The following text has been previously published as:

> Arne Heimendahl, Markus Heinrich and David Gross[1]. "The axiomatic and the operational approaches to resource theories of magic do not coincide". In: Journal of Mathematical Physics 63, 112201 (2022), https://doi.org/10.1063/5.0085774.

Deviations from the journal version are limited to typesetting and notation.These changes were performed to match the rest of this thesis.

Arne Heimendahl and Markus Heinrich are the main contributors to this work and contributed equally. AH constructed the counterexample of a channel that is CSP but not SO and worked out a proof that it is a vertex of the set of CSP maps. Additionally, he contributed to work out the remaining proofs in Section 3.4.

Abstract

Stabilizer operations occupy a prominent role in fault-tolerant quantum computing. They are defined *operationally*: by the use of Clifford gates, Pauli measurements and classical control. These operations can be efficiently simulated on a classical computer, a result which

[1] AH and MH contributed equally.

is known as the *Gottesman-Knill* theorem. However, an additional supply of *magic states* is enough to promote them to a universal, fault-tolerant model for quantum computing. To quantify the needed resources in terms of magic states, a *resource theory of magic* has been developed. stabilizer operations (SO) are considered free within this theory, however they are not the most general class of free operations. From an axiomatic point of view, these are the *completely stabilizer-preserving* (CSP) channels, defined as those that preserve the convex hull of stabilizer states. It has been an open problem to decide whether these two definitions lead to the same class of operations. In this work, we answer this question in the negative, by constructing an explicit counter-example. This indicates that recently proposed stabilizer-based simulation techniques of CSP maps are strictly more powerful than Gottesman-Knill-like methods. The result is analogous to a well-known fact in entanglement theory, namely that there is a gap between the operationally defined class of local operations and classical communication (LOCC) and the axiomatically defined class of separable channels.

3.1 Introduction

Despite the advances in the development of quantum platforms, understanding the precise set of quantum phenomena that is required for a quantum advantage over classical computers remains an elusive task. However, for the design of fault-tolerant quantum computers, it seems imperative to understand these necessary resources. Here, the *magic state model* of quantum computing offers a particularly fruitful perspective. In this model, all operations performed by the quantum computer are divided into two classes. The first class consists of the preparation of stabilizer states, the implementation of Cifford gates, and Pauli measurements. These *stabilizer operations* by themselves can be efficiently simulated classically by the Gottesman-Knill Theorem [Got99, SA04]. Secondly, the quantum computer needs to be able to prepare *magic states*, defined as states that allow for the implementation of any quantum algorithm when acted on by stabilizer

operations [BK05]. In this sense, the magic states provide the "non-classicality" required for a quantum advantage.

During recent years, there has been an increasing interest in developing a resource theory of quantum computing that allows for a precise quantification of *magic*. First resource theories were developed for the somewhat simpler case of odd-dimensional systems, based on a phase-space representation via Wigner functions. There, the total negativity in the Wigner function of a state is a *resource monotone* called *mana*, and non-zero mana is a necessary condition for a quantum speed-up [Gal05, Gro07, VFGE12, VMGE14, ME12, HWVE14, DOBV$^+$17]. In the practically more relevant case of qubits, this theory breaks down, which has led to a number of parallel developments [HC17b, HG19, SC19, RBVT$^+$20, SRP$^+$21, BCHK20, HMMVG21, LW22]. A common element is that the finite set of stabilizer states, or more generally their convex hull, the *stabilizer polytope*, is taken as the set of free states. Since stabilizer operations preserve the stabilizer polytope, they are considered free operations in this theory and any monotones should be non-increasing under those. A number of such *magic monotones* have been studied and their values linked to the runtime of classical simulation algorithms [PWB15, BG16, BBC$^+$19, SC19]. In this sense, the degree of magic present in a quantum circuit does seem to correlate with the quantum advantages it confers – thus validating the premise of the approach.

The set of stabilizer operations (SO) are defined in terms of concrete actions ("prepare a stabilizer state, perform a Clifford unitary, make a measurement, ...") and thus represent an *operational* approach to defining free transformations in a resource theory of magic. It is often fruitful to start from an *axiomatic* point of view, by defining the set of free transformations as those physical maps that preserve the set of free states. This approach has been introduced recently by [SC19]. They suggest to refer to a linear map as *completely stabilizer-preserving* (CSP) if it preserves the stabilizer polytope, even when acting on parts of an entangled system. It has been shown that the magic monotones mentioned above are also non-increasing under CSP maps [SRP$^+$21].

A natural question is therefore whether the two approaches coincide

– i.e. whether SO = CSP, or whether there are CSP maps that cannot be realised as stabilizer operations [SC19].

To build an intuition for the question, consider the analogous problem in entanglement theory, where the free resources are the separable states. The axiomatically defined free transformations are the *separable maps* – completely positive maps that preserve the set of separable states. The operationally defined free transformations are those that can be realised by local operations and classical communication (LOCC). It is known that the set of separable maps is strictly larger than the set of LOCC [BDF$^+$99, CLM$^+$14] – a fact that leads e.g. to a notable gap in the success probability of quantum state discrimination [KTYI07, DFXY09] and entanglement conversion [CCL12] between the two classes.

In this work, we show that – also in resource theories of magic – the axiomatic and the operational approaches lead to different classes, that is SO $\neq$ CSP.

As an auxiliary result, we derive a normal form for stabilizer operations which is used to prove our main result. From this form, it is evident that any stabilizer operation can be realised in a finite number of rounds – a statement which is known to not hold for LOCC operations in entanglement theory [Chi]. Furthermore, we give a characterization of CSP channels in terms of certain generalised stabilizer measurements and adaptive Clifford operations. This characterization has been used in a classical simulation algorithm of CSP channels by [SRP$^+$21].

Outline

In Section 3.2, we give an introduction to the relevant concepts used throughout the main part of this work. Next, we prove a minimal version of our main result and illustrate our proof technique for the 2-qubit case in Section 3.3. There, we show that there is a 2-qubit CSP channel that is not a stabilizer operation. In Section 3.4, we generalise this minimal result to an arbitrary number of qudits. Furthermore, we prove equality of CSP and SO for a single qudit. In Section 3.5, we de-

scribe additional properties of CSP channels and give some examples. We conclude the main part by commenting on potential implications and future work in Sec. 3.6.

3.2 Preliminaries

3.2.1 Stabilizer formalism

Consider the Hilbert space $\mathcal{H} = (\mathbb{C}^d)^{\otimes n}$ of n qudits of dimension d, where we assume that d is prime. We label the computational basis $|x\rangle$ by vectors x in the discrete vector space $\mathbb{F}_d^n$. Here, $\mathbb{F}_d$ is the finite field of d elements which can be taken to be the residue field $\mathbb{Z}/d\mathbb{Z}$ of integers modulo d. Let $\omega = e^{2\pi i/d}$ be a d-th root of unity, then we define the n-qudit Z and X operator as usual by their action on the computational basis:

$$Z(z)\,|y\rangle := \omega^{z\cdot y}\,|y\rangle\,, \quad X(x)\,|y\rangle := |y+x\rangle\,, \quad z,x,y \in \mathbb{F}_d^n. \tag{3.1}$$

Here, all operations take place in the finite field $\mathbb{F}_d$ (i.e. modulo d), if not stated otherwise. To treat the slightly different theory for even and odd d on the same footing, we introduce the convention

$$\tau := (-1)^d e^{i\pi/d}, \qquad D := \begin{cases} 2d & \text{if } d = 2 \\ d & \text{else.} \end{cases} \tag{3.2}$$

Note that τ is always a D-th root of unity such that $\tau^2 = \omega$. We group the Z and X operators and their coordinates to define an arbitrary (generalised) Pauli operator indexed by $a = (a_z, a_x) \in \mathbb{F}_d^{2n}$:

$$w(a) := \tau^{-\gamma(a)} Z(a_z) X(a_x), \qquad \gamma(a) := a_z \cdot a_x \mod D. \tag{3.3}$$

Finally, the *Heisenberg-Weyl* or *generalised Pauli group* is the group generated by Pauli operators and can be written as:

$$\mathcal{P}_n(d) := \langle \{ w(a) \mid a \in \mathbb{F}_d^{2n} \} \rangle = \{ \tau^k w(a) \mid k \in \mathbb{Z}_D, a \in \mathbb{F}_d^{2n} \}. \tag{3.4}$$

The *Clifford group* is defined as the group of unitary symmetries of the Pauli group:

$$\mathrm{Cl}_n(d) := \left\{ U \in U(d^n) \mid U\mathcal{P}_n(d)U^\dagger = \mathcal{P}_n(d) \right\} / U(1). \tag{3.5}$$

We take the quotient with respect to irrelevant global phases in order to render the Clifford group a finite group. If the dimension d is clear from the context, we often omit it to simplify notation.

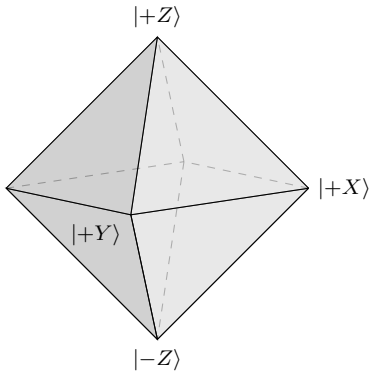


Figure 3.1: Bloch representation of the single-qubit *stabilizer polytope*, which is the octahedron spanned by the six ± 1 eigenstates of the Pauli X,Y, and Z operators. The simple geometry is not representative for the general situation in high dimensions.

An Abelian subgroup $S \subset \mathcal{P}_n(d)$ that does not contain $\omega \mathbb{1}$ is called a *stabilizer group*. The subspace $C(S) \subset (\mathbb{C}^d)^{\otimes n}$ of common fixed points of S is the *stabilizer code* associated with S. One verifies easily that the orthogonal projection onto $C(S)$ is given by

$$P_S = |S|^{-1} \sum_{s \in S} s. \tag{3.6}$$

By taking traces, it follows that the dimension $\dim C(S)$ equals $d^n/|S| = d^{n-k}$, where $k = \operatorname{rank}(S)$ is the rank of S. Hence, S defines a $[[n, n-k]]$ quantum code and we denote by $\mathrm{STAB}(d, n, k)$ the set of these stabilizer codes. Of particular interest is the case $k = n$, for which P_S is rank 1 and thus defines a pure quantum state, called *stabilizer state*. The set of pure stabilizer states $\mathrm{STAB}(d, n) \equiv \mathrm{STAB}(d, n, n)$ spans a convex polytope that is full-dimensional in state space, the *stabilizer polytope* $\mathrm{SP}_n(d) := \operatorname{conv} \mathrm{STAB}(d, n)$. For a single qubit, i.e. $d = 2$ and $n = 1$, this is the well-known octahedron spanned by the Pauli X, Y, Z eigenstates, see Fig. 3.1. Elements of $\mathrm{SP}_n(d)$ will be referred to as *mixed stabilizer states*.

3.2.2 Stabilizer operations

The Gottesman-Knill theorem states that *stabilizer operations* can be simulated in a time which is polynomial in the system size [Got97, SA04]. These operations are defined as follows.

Definition 3.2.1 (stabilizer operation)**.** *A quantum channel taking n input qudits to m output qudits, each of prime dimension d, is a* stabilizer operation*, if it is composed of the following fundamental operations:*

- *preparation of qudits in stabilizer states,*
- *application of Clifford unitaries,*
- *Pauli measurements, and*
- *discarding of qudits.*

An arbitrary random function of previous measurement outcomes can be used to decide which fundamental operation to perform in each step. More precisely, we assume that measurement outcomes are kept until the operation is completed, and are subsequently erased. Hence, the final state is a suitably weighted average over the possible outcome states associated to each set of measurement outcomes. The set of all stabilizer operations is denoted by $\mathrm{SO}_{n,m}(d)$*, with* $\mathrm{SO}_n(d) := \mathrm{SO}_{n,n}(d)$*. If the dimension d is clear from the context, we often omit it to simplify notation.*

Typically, one requires that the classical control logic can be implemented in a computationally efficient way (and the Gottesman-Knill Theorem applies only under this additional assumption). In the present paper we will drop the efficiency requirement and show that even the resulting larger class of stabilizer operations is smaller than the set of CSP channels. As we lay out in Remark 3.2.6, this strengthening of the problem formulation is actually necessary in order to avoid a trivial separation of SO and CSP due to their different computational capabilities.

Because of the possibility to make use of randomness, the set of stabilizer operations $\mathrm{SO}_{n,m}$ is convex. Its extreme points will turn out to play an important role in our construction.

By definition, stabilizer operations can be seen as an iterative protocol where a quantum computer capable of performing fundamental stabilizer operations interacts with a classical control logic. Generalizing results on the structure of Kraus operators of stabilizer operations obtained in Ref. [CB09], we will establish in Thm. 3.4.2 that any operation in $\mathrm{SO}_{n,m}$ requires at most n interactive rounds. This stands in contrast to the class LOCC studied in entanglement theory, where no analogous finite bound exists [Chi].

In our analysis, we will come across the class of stabilizer operations that involve no measurements or classical randomness. This class coincides with the set of channels whose dilation can be realized with a Clifford unitary:

Definition 3.2.2. *A superoperator* $\mathcal{E} : L((\mathbb{C}^d)^{\otimes n}) \to L((\mathbb{C}^d)^{\otimes m})$ *has a* Clifford dilation *if there exists a number* k*, a* k*-qudit stabilizer state* $|s\rangle$*, and a Clifford unitary* U *on* $n+k$ *qudits such that*

$$\mathcal{E}(\rho) = \mathrm{Tr}_{m+1,\ldots,n+k}\left[U(\rho \otimes |s\rangle\langle s|)U^\dagger\right].$$

3.2.3 Completely stabilizer-preserving channels

From a resource-theoretic perspective, the maximal set of free operations is the set of quantum channels which do not generate resources, i. e. which preserve the set of free states, see e. g. Ref. [CG19]. If we take the set of free states to be the *stabilizer polytope* $\mathrm{SP}_n(d)$, the resource non-generating (RNG) channels are the *stabilizer-preserving* (SP) *channels*. For this maximal set of free operations, relatively strong statements can be made from general resource-theoretic arguments. For instance, it has been recently shown that the resource theory with stabilizer-preserving channels is asymptotically reversible which implies that resource-optimal distillation rates can be achieved with stabilizer-preserving channels [LW22].

In general, a resource theory with RNG channels has the disadvantage that it is not closed under tensor products since RNG channels

may fail to be free when applied to subsystems. The class of RNG channels for which this is still the case are the *completely* resource non-generating channels [CG19]. For some resource theories, these two classes coincide, but not for the resource theory of magic [SC19].

Following this idea, [SC19, SRP+21] have studied completely stabilizer-preserving (CSP) channels as the free operations in a resource theory of magic state quantum computing.

Definition 3.2.3. *A superoperator* $\mathcal{E} : L((\mathbb{C}^d)^{\otimes n}) \to L((\mathbb{C}^d)^{\otimes m})$ *is called* completely stabilizer-preserving *(CSP) if and only if* $\mathcal{E}\otimes\mathrm{id}_k(\mathrm{SP}_{n+k}(d)) \subset \mathrm{SP}_{m+k}(d)$ *for all* $k \in \mathbb{N}$*. The set of CSP maps is denoted by* $\mathrm{CSP}_{n,m}(d)$ *and* $\mathrm{CSP}_n(d) := \mathrm{CSP}_{n,n}(d)$*. If the dimension* d *is clear from the context, we often omit it to simplify notation.*

As it is the case for completely positive maps, one can show that it is indeed enough to check the condition for $k = n$ [SC19, Lem. 4.1].

It will be helpful to characterise CSP maps via their *Choi-Jamiołkowski representation*. Recall that in this representation, a linear map $\mathcal{E} : L((\mathbb{C}^d)^{\otimes n}) \to L((\mathbb{C}^d)^{\otimes m})$ is associated with an operator

$$\mathcal{J}(\mathcal{E}) := \mathcal{E} \otimes \mathrm{id}_n(\left|\phi^+\right\rangle\!\left\langle\phi^+\right|) \in L((\mathbb{C}^d)^{\otimes m}) \otimes L((\mathbb{C}^d)^{\otimes n}), \tag{3.7}$$

where $\left|\phi^+\right\rangle = d^{-n} \sum_{x\in\mathbb{F}_d^n} |xx\rangle$ is the standard maximally entangled state with respect to the computational basis. Choi's theorem states that $\mathcal{E}$ is completely positive if and only if its Choi-Jamiołkowski representation lies in the positive semidefinite cone

$$\mathrm{PSD}_{n+m}(d) \subset L((\mathbb{C}^d)^{\otimes m}) \otimes L((\mathbb{C}^d)^{\otimes n}). \tag{3.8}$$

What is more, the map $\mathcal{E}$ is trace-preserving if and only if its Choi-Jamiołkowski representation lies in the affine space

$$\mathrm{TP}_{n,m}(d) = \left\{\rho \in L((\mathbb{C}^d)^{\otimes m}) \otimes L((\mathbb{C}^d)^{\otimes n}) \mid \mathrm{Tr}_1\, \rho = \mathbb{1}/d^m\right\}. \tag{3.9}$$

In particular, for the set $\mathrm{CPTP}_{n,m}(d)$ of completely positive and trace-preserving maps, we have the characterization

$$\mathcal{J}(\mathrm{CPTP}_{n,m}(d)) = \mathrm{PSD}_{n+m}(d) \cap \mathrm{TP}_{n,m}(d). \tag{3.10}$$

We now turn to the CSP version of this theory. It turns out that the CSP property has strong implications:

Lemma 3.2.4. *Any CSP map is completely positive and trace-preserving.*

Proof. The first claim follows from the Choi-Jamiołkowski Theorem, because $|\phi^+\rangle$ is a stabilizer state. As for the second claim: Because the set of stabilizer states (as projections) spans $L((\mathbb{C}^d)^{\otimes n})$, every Hermitian trace-one operator can be written as an affine combination of stabilizer states. By definition, any CSP map maps this to an affine combination of stabilizer states in the output space $L((\mathbb{C}^d)^{\otimes m})$. In particular, it is trace-preserving. □

The CSP-analogue of Eq. (3.10) was proven in Ref. [SC19].

Lemma 3.2.5 (Lem. 4.2 in [SC19])**.** *A linear map $\mathcal{E} : L((\mathbb{C}^d)^{\otimes n}) \to L((\mathbb{C}^d)^{\otimes m})$ is CSP if and only if its Choi representation lies in the intersection of the stabilizer polytope with the affine space* $\mathrm{TP}_{n,m}(d)$*:*

$$\mathcal{J}(\mathrm{CSP}_{n,m}(d)) = \mathrm{SP}_{n+m}(d) \cap \mathrm{TP}_{n,m}(d). \tag{3.11}$$

In particular, $\mathrm{CSP}_{n,m}(d)$ *is a convex polytope.*

Additional properties of CSP channels, as well as a collection of examples, are provided in Sec. 3.5.

For this work, the focus lies on channels which map the input space to itself, i.e. $n = m$. In the main part of this paper, we study the relation between completely stabilizer-preserving channels $\mathrm{CSP}_n(d)$ and stabilizer operations $\mathrm{SO}_n(d)$. In particular, we show that they agree if and only if $n = 1$. The definitions in this section, as well as the general version of our main result, apply both to qubits $d = 2$ and to qudits, where d is an odd prime number.

However, we point out that in odd prime dimensions, the set of free states can be enlarged to include all states with a non-negative Wigner function. This is a convex set $\mathcal{W}_n^+(d)$ given as the intersection of a probability simplex with the cone of positive-semidefinite matrices, and strictly larger than the stabilizer polytope [Gro06, Gro07]. The resulting resource theory differs quite significantly from the qubit case [VFGE12, VMGE14, ME12] and naturally leads to a different class of resource-non generating channels, namely those which do not

induce Wigner negativity, see e. g. Ref. [WWS19]. Thus, the questions we ask are arguably better motivated in the qubit case.

Another difference between the resource theories in even and odd dimensions is given by [ADGS18]. They show that for a single qutrit, there is a stabilizer-preserving channel which can induce negativity in a state's Wigner function, in particular it cannot be a stabilizer operation. This shows that SP channels are not the correct free operations for a resource theory of magic in odd dimensions. In contrast, we show in this work that the set of *completely* stabilizer-preserving channels agrees with the set of stabilizer operations for a single qudit, independent of the dimension. Moreover, arbitrary multi-qudit CSP channels for odd d cannot induce negativity in the Wigner function by the following argument. Analogous to Lemma 3.2.5, one can show that the set $\mathrm{CWPP}_{n,m}(d)$ of completely $\mathcal{W}_n^+(d)$-preserving channels corresponds to $\mathcal{W}_{n+m}^+(d) \cap \mathrm{TP}_{n,m}(d)$. This follows from the Choi-Jamiołkowski inversion formula and the fact that $|\phi^+\rangle\langle\phi^+| \in \mathrm{SP}_{n+m}(d) \subset \mathcal{W}_{n+m}^+(d)$. Therefore, $\mathrm{CSP}_{n,m}(d)$ is contained in $\mathrm{CWPP}_{n,m}(d)$ and, in particular, cannot induce negativity in the Wigner function. This establishes the chain of inclusions $\mathrm{SO}_{n,m}(d) \subset \mathrm{CSP}_{n,m}(d) \subset \mathrm{CWPP}_{n,m}(d)$ for odd prime d, where our main result 6.3.1 implies that the first inclusion is proper for $n, m > 1$. While one cannot readily dismiss the possibility that the last inclusion is an equality, we conjecture that it is indeed proper, too.

Remark 3.2.6. *The definition 3.2.3 of CSP allows for quantum channels which are of the form [Cam21]*

$$\mathcal{E}(|x\rangle\langle y|) := \delta_{x,y}\, |O(x)\rangle\langle O(x)| \tag{3.12}$$

where O can be an arbitrary Boolean function. The definition does not preclude one to consider families $\mathcal{E}_n$ of channels that are associated with Boolean functions O_n that are not Turing computable (e.g. functions that decide the halting problem). The discussion shows that it is meaningless to compare stabilizer operations with *computational efficiency requirements to CSP channels defined* without *such constraints. To avoid a trivial separation of the classes, we show here that*

even stabilizer operations where the classical control logic can consist of arbitrary random functions of previous measurement results cannot implement all CSP channels.

3.3 The CSP class is strictly larger than the class of stabilizer operations

In this section, we prove a minimal version of the main result. The general version, treating the multi-qudit case, is stated and proven in Sec. 3.4.

Theorem 3.3.1. *For two qubits, the set* $\mathrm{CSP}_2(2)$ *is strictly larger than* $\mathrm{SO}_2(2)$.

Concretely, we will establish that the following two-qubit channel is completely stabilizer-preserving, but not a stabilizer operation:

$$\Lambda(\rho) := \rho_{00,00} \left|{+}{+}\right\rangle\!\left\langle{+}{+}\right| + \sum_{x\in\{01,10,11\}} \rho_{x,x} \left|x\right\rangle\!\left\langle x\right| + \frac{1}{2} \sum_{\substack{x,y\in\{01,10,11\}\\ x\neq y}} \rho_{x,y} \left|x\right\rangle\!\left\langle y\right|, \tag{3.13}$$

where $|+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$ and $\rho_{x,y} = \langle x|\rho|y\rangle$.

The intuition behind the counter-example is as follows: First, consider a projective measurement that distinguishes between $|00\rangle$ and its orthocomplement. It is plausible that one cannot implement such a measurement using stabilizer operations – if for no other reason than that Pauli measurements lead to Kraus operators whose rank is a power of two. The channel Λ may be realized by such a measurement, followed by the application of Hadamard gates on all qubits when the outcome $|00\rangle$ is obtained, or a partial dephasing operation in the alternate case. It turns out that the second step makes Λ CSP, while the no-go argument concerning the measurement remains valid.

Appendix 3.B describes some properties of Λ that are not directly required for the proof below.

In the proof, we will use the fact that the channel (3.13) is an extreme point in the convex set $\mathrm{CSP}_2 \equiv \mathrm{CSP}_2(2)$. To show this, it

turns out to be sufficient to restrict attention to the intersection of CSP_2 with a fairly low-dimensional affine space – a step that greatly simplifies the description of the convex geometry.

Concretely, we define the convex set of *almost-diagonal channels* AD_2 as the set of two-qubit quantum channels $\mathcal{E}$ that act on the pure states of the computational basis in the following way:

$$\mathcal{E}(|00\rangle\langle 00|) = |{+}{+}\rangle\langle{+}{+}|\,, \qquad \mathcal{E}(|x\rangle\langle x|) = |x\rangle\langle x| \quad x \in \{01, 10, 11\}. \tag{3.14}$$

By comparison with Eq. (3.13) it is immediate that Λ lies in $\mathrm{CSP}_2 \cap \mathrm{AD}_2$. This intersection is isomorphic, as a convex set, to a subpolytope of the two qubit stabilizer polytope.

Definition 3.3.2. *Let* P_2 *be the polytope of complex* 4×4 *matrices* σ *that (1) are a convex combination of two-qubit stabilizer states, and (2), when expressed in the basis* $\{|00\rangle, |01\rangle, |10\rangle, |11\rangle\}$*, are of the form*

$$\sigma = \frac{1}{3}\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & * & * \\ 0 & * & 1 & * \\ 0 & * & * & 1 \end{pmatrix}$$

with $*$*'s denoting arbitrary complex values.*

Lemma 3.3.3. *A map* $\mathcal{E}$ *lies in* $\mathrm{CSP}_2 \cap \mathrm{AD}_2$ *if and only if there exists a* $\sigma \in \mathrm{P}_2$ *such that*

$$\mathcal{E}(\rho) = 3\,\sigma \circ \rho + \langle 00|\rho|00\rangle\ |{+}{+}\rangle\langle{+}{+}|\,, \tag{3.15}$$

where $\circ$ *is the* Hadamard *(or element-wise) product. In particular, the polytopes* $\mathrm{CSP}_2 \cap \mathrm{AD}_2$ *and* P_2 *are isomorphic.*

Proof. **"Only if":** Assume that $\mathcal{E}$ is CSP, i.e. its Choi state is expressible as

$$\mathcal{J}(\mathcal{E}) = \sum_{s \in \mathrm{STAB}(4)} p_s |s\rangle\langle s|. \tag{3.16}$$

The Choi state has the property that

$$\mathcal{E}(|x\rangle\langle y|) = 4(\mathbb{1} \otimes \langle x|)\, \mathcal{J}(\mathcal{E})\, (\mathbb{1} \otimes |y\rangle) = 4 \sum_s p_s\, (\mathbb{1} \otimes \langle x|)|s\rangle\, \langle s|(\mathbb{1} \otimes |y\rangle) \quad \forall x, y \in \mathbb{F}_2^2 \tag{3.17}$$

Evaluating Eq. (3.17) on the diagonal and using Eq. (3.14) implies that, for all s with $p_s \neq 0$,

$$(\mathbb{1} \otimes \langle 00|)|s\rangle \propto |{+}{+}\rangle\,, \tag{3.18}$$

$$(\mathbb{1} \otimes \langle x|)|s\rangle \propto |x\rangle \quad \forall x \neq 00, \tag{3.19}$$

where $\propto$ denotes equality up to a proportionality constant including 0.

There must be at least one $|s\rangle$ with $(\mathbb{1} \otimes \langle 00|)|s\rangle \neq 0$. We claim that this implies $|s\rangle = |{+}{+}\rangle|00\rangle$ and $p_s = \frac{1}{4}$. Indeed, assume for the sake of reaching a contradiction that $|s\rangle$ has Schmidt rank larger than one. Then for at least one $x \in \mathbb{F}_2^2$, the contraction $(\mathbb{1} \otimes \langle x|)|s\rangle$ is not proportional to $|{+}{+}\rangle$. By a well-known property of stabilizer states (c.f. Prop. 3.A.2), $(\mathbb{1} \otimes \langle x|)|s\rangle$ is then orthogonal to $|+-\rangle$, which contradicts (3.19). Thus $|s\rangle$ is a product state. The claimed form follows from (3.18), and the value of p_s from (3.17).

We now treat the terms $|s\rangle$ different from $|{+}{+}\rangle\,|00\rangle$. Equations (3.18, 3.19) and Proposition 3.A.2 imply that these stabilizer states are "diagonal in the computational basis" in the sense that

$$|s\rangle = \sum_{x \in \mathbb{F}_2^2} \tilde{s}(x)|x\rangle|x\rangle \qquad \text{for some } \tilde{s} : \mathbb{F}_2^2 \to \mathbb{C} \text{ with } \tilde{s}(00) = 0.$$

Define the n-qudit state $|\tilde{s}\rangle = \sum_x \tilde{s}(x)\,|x\rangle$. Then $|\tilde{s}\rangle$ is orthogonal to $|00\rangle$. It is also a normalised stabilizer state, because it arises from the action of a Clifford unitary on $|s\rangle$:

$$|\tilde{s}\rangle \otimes |00\rangle = CX_{1,3}\, CX_{2,4}\, |s\rangle\,,$$

where $CX_{i,j}$ is the controlled-NOT gate with the i-th qubit controlling the j-th one. Setting

$$\sigma = \frac{4}{3} \sum_{s \neq |{+}{+}\rangle|00\rangle} p_s\, |\tilde{s}\rangle\langle\tilde{s}| \in \mathrm{P}_2,$$

we get that for all $(x, y) \neq (00, 00)$,

$$\begin{aligned}\mathcal{E}(|x\rangle\langle y|) &= 4\sum_s p_s \left(\mathbb{1} \otimes \langle x|\right) |s\rangle\langle s| \left(\mathbb{1} \otimes |y\rangle\right) \\ &= 4 \sum_{s \neq |{+}{+}\rangle\,|00\rangle} p_s\, \tilde{s}(x)\overline{\tilde{s}(y)}\; |x\rangle\langle y| \\ &= 3\,\sigma \circ |x\rangle\langle y|\,.\end{aligned}$$

"If": The construction above can be reversed straight-forwardly. □

Under the correspondence given in Lemma 3.3.3, the channel Λ defined in Eq. (3.13) corresponds to the matrix

$$\lambda = \frac{1}{6}\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 2 & 1 & 1 \\ 0 & 1 & 2 & 1 \\ 0 & 1 & 1 & 2 \end{pmatrix}. \tag{3.20}$$

Using the relative simplicity of the polytope P_2, we can now show that Λ is an extremal CSP channel.

Lemma 3.3.4. *The matrix λ in Eq. (3.20) is a vertex of P_2. What is more, Λ is a vertex of CSP_2.*

Proof. We will establish the first claim by showing that λ is the unique maximizer in P_2 of the linear functional

$$L : \mathrm{P}_2 \to \mathbb{R}, \qquad \sigma \mapsto \langle +|\sigma|+\rangle = \sum_s p_s |\langle +|s\rangle|^2.$$

There are 15 stabilizer states $|s\rangle$ orthogonal to $|00\rangle$, given by

$$|01\rangle\,, \qquad 2^{-1/2}\big(\,|01\rangle + \omega\,|10\rangle\,\big), \quad \omega \in \{1, -1, i, -i\},$$

and their images under permutations of $\{|01\rangle\,, |10\rangle\,, |11\rangle\}$. Among those, the inner product $|\langle +|s\rangle|^2$ attains its maximum (of $1/2$) exactly for the three cases $2^{-1/2}\big(\,|01\rangle + |10\rangle\,\big), 2^{-1/2}\big(\,|01\rangle + |11\rangle\,\big), 2^{-1/2}\big(\,|10\rangle + |11\rangle\,\big)$. Among the linear combinations of their projection operators, a uniform mixture is the unique solution to the three constraints $\sigma_{x,x} = 1/3$. This solution is equal to λ.

To prove the second claim, assume that $\mathrm{AD}_2 \cap \mathrm{CSP}_2 \ni \mathcal{E} = p\mathcal{E}_1 + (1-p)\mathcal{E}_2$ for some CSP maps $\mathcal{E}_1$ and $\mathcal{E}_2$, and $p \in [0,1]$. The extremality of the pure states on the right hand sides of Eq. (3.14) forces $\mathcal{E}_1$ and $\mathcal{E}_2$ to fulfil the same constraints, i.e. $\mathcal{E}_1, \mathcal{E}_2 \in \mathrm{AD}_2 \cap \mathrm{CSP}_2$. Hence, a channel $\mathcal{E} \in \mathrm{AD}_2 \cap \mathrm{CSP}_2$ is extremal in CSP_2 if and only if it is extremal in the subpolytope $\mathrm{AD}_2 \cap \mathrm{CSP}_2$. □

If Λ was a stabilizer operation, Lem. 3.3.4 would imply that it is extremal in the convex set SO_2. This is because extremality of a point in a convex set implies extremality in every convex subset containing the point. Our strategy now is to identify a property shared by all extremal stabilizer operations, and then to show that Λ fails to posses it.

Theorem 3.3.5 (Pauli invariance of extremal stabilizer operations)**.** *Let $\mathcal{O} \in \mathrm{SO}_2$ be an extremal stabilizer operation that does not have a Clifford dilation. Then the kernel of $\mathcal{O}$ contains a Pauli operator.*

The proof will make use of the following lemma. It says that the operation "preparing an ancilla stabilizer state and performing a Pauli measurement jointly on an input and the ancilla" can be replaced by a random Clifford channel, if the stabilizer state is not an eigenstate of the Pauli operator. (One could also approach the statement through the theory of quantum error correction. In this language, the measured Pauli is a correctable error for the stabilizer code $(\mathbb{C}^2)^{\otimes n} \otimes |s\rangle$, and the Clifford unitaries that appear are the ones correcting the projections onto the eigenspaces of the Pauli operator.)

Lemma 3.3.6. *Let $w(a) \otimes w(b)$ be an $(n+k)$-qubit Pauli operator. Denote the projectors onto the two eigenspaces of $w(a) \otimes w(b)$ by $P_\pm$. Let $|s\rangle$ be a k-qubit stabilizer state that is* not *an eigenstate of $w(b)$. Then there are two $(n+k)$-qubit Clifford unitaries $U_\pm$ such that, for all n-qubit states $|\psi\rangle$, we have $P_\pm |\psi\rangle \otimes |s\rangle = \frac{1}{\sqrt{2}} U_\pm |\psi\rangle \otimes |s\rangle$.*

Proof. There is a k-qubit Clifford unitary V which maps $|s\rangle \mapsto |0^k\rangle$ and $(w(b)|s\rangle) \mapsto |1\rangle|0^{k-1}\rangle$. There is also an n-qubit Clifford U such that $Uw(a)U^\dagger = Z_1$. It thus suffices to show the claim for the special

case $w(a) = Z_1$ and $w(b)|0^k\rangle = |1\rangle|0^{k-1}\rangle$. In terms of a controlled Z-gate $CZ_{(n+1),1}$ (first ancilla qubit controlling the first input qubit):

$$P_{\pm}(|\psi\rangle\otimes|0^k\rangle) = \frac{1}{2}\Big[\ |\psi\rangle\otimes|0\rangle \pm (Z_1\,|\psi\rangle)\otimes|1\rangle\ \Big]\otimes|0^{k-1}\rangle = \frac{1}{\sqrt{2}}\Big[CZ_{(n+1),1}\,|\psi\rangle\ |\pm\rangle\ \Big]\otimes|0^{k-} \tag{3.21}$$

Hence, we can choose $U_+ = CZ_{(n+1),1}H_{n+1}$ and $U_- = CZ_{(n+1),1}H_{n+1}X_{n+1}$ where H_{n+1} and X_{n+1} are the Hadamard and X gate acting on the first ancilla qubit, respectively. □

Proof of Theorem 3.3.5. Consider an implementation of $\mathcal{O}$ using elementary Clifford operations. By extremality, we may assume that no classical randomness is used. Thus the implementation must contain at least one Pauli measurement, for else $\mathcal{O}$ would have a Clifford dilation. Propagating the first Pauli measurement past preceding Clifford unitaries if necessary, there is no loss of generality in assuming that the implementation starts by preparing k ancilla qubits in a stabilizer state $|s\rangle$ and then immediately measures an $(n+k)$-qubit Pauli operator $w(a)\otimes w(b)$ with $a \in \mathbb{F}_2^{2n}$ and $b \in \mathbb{F}_2^{2k}$.

We will show now that one may in fact assume that $a \neq 0$ and $b = 0$, i.e. that the implementation starts by measuring a non-trivial Pauli without involving the ancillas.

Indeed, if $a = 0$, the measurement only acts on the ancilla systems. We can thus write $\mathcal{O} = p_1\mathcal{O}_1 + p_{-1}\mathcal{O}_{-1}$, where $\mathcal{O}_\pm$ are the operations conditioned on the outcome, and the probabilities $p_\pm$ do not depend on the input state. Extremality implies that either $\mathcal{O}_1 = \mathcal{O}_{-1}$ or only one of the $p_\pm$ differs from 0. Hence one can eliminate the measurement from the implementation and restart the proof. Iterating this argument if necessary, we will eventually obtain a Pauli measurement with $a \neq 0$, as $\mathcal{O}$ does not have a Clifford dilation.

Next assume that $b \neq 0$. First consider the case where $|s\rangle$ is not an eigenstate of $w(b)$. By Lemma 3.3.6, the measurement can be replaced by a process that applies one of two Clifford unitaries, each with probability 1/2. Arguing as above, this process either contradicts extremality or can be eliminated. Thus we may assume that $|s\rangle$ is an eigenstate of $w(b)$. In this case, the ancilla system affects

the measurement process only by changing the labels of the measurement results (specifically by multiplying them with the eigenvalue). Absorbing this deterministic relabelling into any classical control, we may set $b = 0$.

Let $P_\pm = \frac{1}{2}(\mathbb{1} \pm w(a))$ be the projections onto the eigenspaces of $w(a)$. Choose any two-qubit Pauli operator $w(u)$ that anti-commutes with $w(a)$. Using the above expression for $P_\pm$, one finds $P_\pm w(u) P_\pm = 0$ and thus $w(u)$ is in the kernel of both initial branches of $\mathcal{O}$, hence $w(u) \in \ker \mathcal{O}$. □

The following lemma is thus sufficient to establish Theorem 3.3.1.

Lemma 3.3.7. *Let* Λ *be as in Eq.* (3.13)*. Then* Λ *has no Clifford dilation, and* $\ker \Lambda$ *does not contain a Pauli operator.*

Proof. Assume, for the sake of reaching a contradiction, that Λ does have a Clifford dilation. Then $\Lambda^\dagger$ maps Pauli operators to Pauli operators, up to a phase. From Eq. (3.13):

$$0 = \langle ++| \, Z_1 \, |++\rangle = \mathrm{Tr}\left(|00\rangle\langle 00| \, \Lambda^\dagger(Z_1)\right), \tag{3.22}$$

$$(-1)^x = \langle xy| \, Z_1 \, |xy\rangle = \mathrm{Tr}\left(|xy\rangle\langle xy| \, \Lambda^\dagger(Z_1)\right), \quad |xy\rangle \in \{|01\rangle, |10\rangle, |11\rangle\}. \tag{3.23}$$

Eq. (3.22) implies that $\Lambda^\dagger(Z_1)$ is proportional to X or Y on at least one of the factors. This, however, is incompatible with Eq. (3.23), which is the sought-for contradiction.

One reads off Eq. (3.13) that $\Lambda(|x\rangle\langle y|) = 0$ if and only if $x = 0$ and $y \neq 0$ or $x \neq 0$ and $y = 0$. That means that the kernel of Λ consists of the operators of the form

$$\begin{pmatrix} 0 & * & * & * \\ * & 0 & 0 & 0 \\ * & 0 & 0 & 0 \\ * & 0 & 0 & 0 \end{pmatrix}, \tag{3.24}$$

each of which has rank at most 2. In particular, this rules out Pauli operators. □

3.4 General formulation

In this section, we generalise Theorem 3.3.1 to our main result: $\mathrm{CSP}_n(d)$ strictly contains $\mathrm{SO}_n(d)$ for any (prime) dimension d and system size $n \geq 2$.

Theorem 3.4.1 ($\mathrm{SO}_n \subsetneq \mathrm{CSP}_n$)**.** *For any prime dimension d, we have $\mathrm{CSP}_n(d) = \mathrm{SO}_n(d)$ if and only if $n = 1$. In particular, the set of CSP maps is strictly larger than the set of stabilizer operations for $n \geq 2$.*

The proof of Theorem 6.3.1 is accomplished in two parts. The equality in the case $n = 1$ is proven independently in Sec. 3.4.3. For the case $n \geq 2$, we start with an identical approach as in Sec. 3.3 and concentrate on the intersection of $\mathrm{CSP}_n(d)$ with almost-diagonal (AD) channels. Although the proof strategy of Sec. 3.3 based on Pauli invariances, i.e. Theorem 3.3.5, also works for arbitrary d and $n \geq 2$, we follow a more direct route in this section. As we show, the restriction to AD channels directly simplifies the description of both general CSP channels and stabilizer operations considerably. Using this result, it is then straightforward to define a linear functional L which separates the almost-diagonal CSP channels from stabilizer operations. As we show, this linear functional is again maximal on a generalisation of the Λ channel to be defined later, cp. Eq. (3.13).

To arrive at the mentioned simplification for stabilizer operations, we first derive a suitable "normal form" in Sec. 3.4.1.

3.4.1 Normal form for stabilizer operations

In this section, we show that any stabilizer operation is a convex combination of circuits performing a projective stabilizer measurement on the input followed by a global, ancilla-assisted Clifford unitary conditioned on the measurement outcome.

Theorem 3.4.2 (Kraus decomposition of SO)**.** *Consider the family of stabilizer operations in $\mathrm{SO}_{n,m}(d)$ of the following type:*

$$\mathcal{E}(\rho) = \mathrm{Tr}_{m+1,\ldots,n+r} \sum_i U_i \left(P_i \rho P_i \otimes |0^r\rangle\langle 0^r|\right) U_i^\dagger, \tag{3.25}$$

where $\{P_i\}$ is a projective measurement given by mutually orthogonal stabilizer code projectors and the U_i's are Clifford unitaries acting on $n+r$ qudits. Then, the following holds:

(i) Any $\mathcal{O} \in \mathrm{SO}_{n,m}(d)$ is a convex combination of SO of the above type (3.25)*.*

(ii) In particular, any stabilizer operation can be realised in at most n rounds.

Remark 3.4.3. *A projective measurement composed of mutually orthogonal stabilizer code projectors is not necessarily associated to a single set of mutually commuting Pauli operators (i.e. a syndrome measurement). An example for this is the measurement of the basis $\{|00\rangle, |01\rangle, |1+\rangle, |1-\rangle\}$.*

The proof of Theorem 3.4.2 is similar to related results in Ref. [CB09] and Ref. [BCHK20, Thm. 5.3]. However, the latter works focus on the form of *post-selected* stabilizer operations, i.e. on the form of a single Kraus operator in Eq. (3.25). Moreover, Ref. [BCHK20] only considers the form of post-selected stabilizer operations which map a fixed input to a fixed output state. Here, we show that a careful argumentation allows us to manipulate all Kraus operators simultaneously to arrive at a similar result for the entire quantum channel. The

To prove Theorem 3.4.2, we use Lemmata 3.4.4 and 3.4.5 to eliminate non-commuting Pauli measurements and Pauli measurements on ancilla qudits. In this way, an arbitrary stabilizer operation can be iteratively decomposed into a convex combination of stabilizer operations of the form (3.25).

Lemma 3.4.4 generalises [CB09, Sec. 6] and [BCHK20, Prop. A8] to arbitrary prime dimension d.

Lemma 3.4.4. *Suppose P_1 and P_2 are non-commuting $[[n, n-1]]$ stabilizer code projectors. Then, $P_1 P_2 = d^{-1/2} V P_2$ for a suitable Clifford unitary V.*

Proof. Pairs of non-commuting $[[n, n-1]]$ stabilizer codes form a single orbit under the Clifford group. To see this, let w_1 and w_2 be Pauli

operators that generate such a pair P_1 and P_2 and let $\tilde{w}_1$, $\tilde{w}_2$ generate another pair $\tilde{P}_1$ and $\tilde{P}_2$. By redefining the generators with a suitable power of ω, we can assume that $w_1 w_2 = \omega w_2 w_1$ and $\tilde{w}_1 \tilde{w}_2 = \omega \tilde{w}_2 \tilde{w}_1$. Then there is a Clifford unitary U mapping w_1 to $\tilde{w}_1$ and w_2 to $\tilde{w}_2$ and thus P_1 to $\tilde{P}_1$ and P_2 to $\tilde{P}_2$ as claimed. Thus, we may assume that $P_1 = |+\rangle\langle+| \otimes \mathbb{1}_{n-1}$ and $P_2 = |0\rangle\langle 0| \otimes \mathbb{1}_{n-1}$ and clearly $P_1 P_2 = |+\rangle\langle+|\,|0\rangle\langle 0| \otimes \mathbb{1}_{n-1} = \frac{1}{\sqrt{d}} H P_2$ where H is the Hadamard gate. □

The following lemma is a generalisation of Lemma 3.3.6 to any prime dimension d.

Lemma 3.4.5. *Let $w(a) \otimes w(b)$ be a $(n+k)$-qudit Pauli operator and let $|s\rangle$ be a k-qudit stabilizer state which is not an eigenstate of $w(b)$. For any $x \in \mathbb{F}_d$, denote the projector onto the eigenspace of $w(a) \otimes w(b)$ with eigenvalue ω^x by P_x. Then, there are Clifford unitaries U_x such that $P_x\,|\psi\rangle \otimes |s\rangle = d^{-1/2} U_x\,|\psi\rangle \otimes |s\rangle$ for all $\psi \in (\mathbb{C}^d)^{\otimes n}$ and $x \in \mathbb{F}_d$.*

Proof. Since $|s\rangle$ is not an eigenstate of $w(b)$, the stabilizer states $w(xb)\,|s\rangle$ for $x \in \mathbb{F}_d$ are part of the same stabilizer basis. In particular, there is a Clifford unitary V such that $V w(xb)\,|s\rangle = |x\rangle\,|0^{k-1}\rangle$ Moreover, there is a Clifford unitary U such that $U w(a) U^\dagger = Z_1$. Thus, up to acting with U on the input register, and with V on the ancilla register, we may assume that $w(a) = Z_1$ and $w(xb)|0^k\rangle = |x\rangle|0^{k-1}\rangle$. In terms of a controlled Z-gate $CZ_{n+1,1}$ (first ancilla qudit controlling the first input qudit), the action of the projections onto the eigenspaces of $w(a) \otimes w(b)$ is then given by

$$P_x\,|\psi\rangle\otimes|0^k\rangle = \frac{1}{d}\Big[\sum_{y\in\mathbb{F}_d} \omega^{xy}\big(Z_1^x\,|\psi\rangle\big)\otimes|x\rangle\Big]\otimes|0^{k-1}\rangle = \frac{1}{\sqrt{d}}\Big[CZ_{n+1,1}\big(\,|\psi\rangle\otimes H\,|x\rangle\,\big)\Big]\otimes|0^{k-} \tag{3.26}$$

Thus, the claim holds for the Clifford unitary $U_x := CZ_{n+1,1} H_{n+1} X_{n+1}(x)$. □

Proof of Theorem 3.4.2. Suppose $\mathcal{O}$ is a stabilizer operation which does not explicitly use classical randomness and involves l Pauli measurements with outcomes labelled by the ditstring $x = (x_1, \ldots, x_l) \in \mathbb{F}_d^l$. Let us introduce the shorthand notation $x_{[k]} := (x_1, \ldots, x_k)$. Since the partial trace is linear and the size of the ancilla system stays fixed

throughout the proof, we can ignore the possibility of tracing out qudits. Hence, $\mathcal{O}$ can be taken as follows:

$$\mathcal{O}(\rho) = \sum_{x\in\mathbb{F}_d^l} K(x)\rho\otimes|0^r\rangle\langle 0^r|K(x)^\dagger, \quad K(x) = U(x)P(x_l\,|\,x_{[l-1]})P(x_{l-1}\,|\,x_{[l-2]})\cdots P(x_1 \tag{3.27}$$

Without loss of generality, the Kraus operators $K(x)$ are given by consecutive projectors P associated to outcomes of Pauli measurements, and a global Clifford unitary U at the end. All operations may be conditioned on previous measurement outcomes. The projectors fulfil the POVM condition $\sum_{x_k} P(x_k\,|\,x_{[k-1]}) = \mathbb{1}$ for all k and previous outcomes $x_{[k-1]} \in \mathbb{F}_d^{k-1}$. We can visualise the SO as a regular tree with root given by the initial measurement and branches corresponding to sequences of measurement outcomes. The vertices of the tree are labelled by Pauli measurements (see Fig. 3.2).

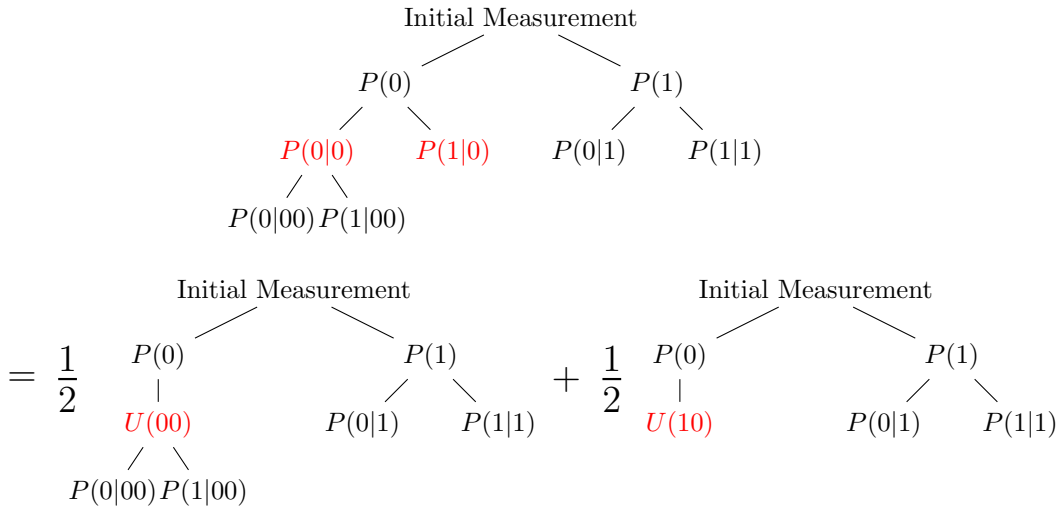


Figure 3.2: Illustration of a tree model of a qubit stabilizer operation. The nodes correspond to outcomes of measurements which in turn depend on the previous outcomes $x_{[k]} \in \mathbb{F}_2^k$. We omit all nodes given by trivial measurements. If a Pauli measurement acts non-trivially on the ancilla state (in the given case $P(0|0)$ and $P(1|0)$ in red), the SO coincides with a uniform convex combination of two SOs, where the measurements $P(0|0)$ and $P(1|0)$ are replaced by Clifford unitaries $U(00)$ and $U(10)$.

We first argue that we can write $\mathcal{O}$ as a convex combination of stabilizer operations which do not measure ancilla qudits. To this end, we

use Lemma 3.4.5 to replace any Pauli measurement involving the ancilla system by a convex combination of suitable Clifford unitaries acting on input and ancilla system. We prove this via induction over the depth k of the tree, starting from the root $k = 1$ and progressing to the leaves $k = l$. Assume that up to depth $k-1$, all measurements are acting trivially on the ancilla system. This implies that in every branch, the ancilla system is still in the initial state $|0^r\rangle$. Let X_{k-1} be the set of previous outcomes $x_{[k-1]}$, such that the k-th measurement conditioned on $x_{[k-1]} \in X_{k-1}$ acts non-trivially on ancilla qudits. By Lemma 3.3.6, we can then write $P(x_k|x_{[k-1]})\,|\psi\rangle \otimes |0^r\rangle = d^{-1/2}U(x_{[k]})\,|\psi\rangle \otimes |0^r\rangle$ for all input states $|\psi\rangle$ and suitable Clifford unitaries $U(x_{[k]})$. For branches starting with $x_{[k-1]}$, we can thus treat x_k as the outcome of a classical, uniformly distributed random variable Y. By conditioning on the outcome $y \in \mathbb{F}_d$ of this random variable, we get new stabilizer operations $\mathcal{O}(y)$ given by the Kraus operators

$$K'(x_l, \ldots, x_{k+1}, y, x_{[k-1]}) := d^{1/2} K(x_l, \ldots, x_{k+1}, y, x_{[k-1]}), \quad x_{[k-1]} \in X_{k-1}, \tag{3.28}$$

$$K'(x_l, \ldots, x_k, x_{[k-1]}) := K(x_l, \ldots, x_k, x_{[k-1]}), \qquad x_{[k-1]} \notin X_{k-1}. \tag{3.29}$$

$\mathcal{O}(y)$ performs the same operation as $\mathcal{O}$ if the first $k-1$ outcomes are not in X_{k-1} and otherwise applies the Clifford unitary $U(y, x_{[k-1]})$ and follows the branch determined by $(y, x_{[k-1]})$, see Fig. 3.2. In particular, it is indeed a stabilizer operation and $\mathcal{O} = d^{-1} \sum_y \mathcal{O}(y)$. Moreover, all measurements in $\mathcal{O}(y)$ act trivially on the ancilla system up to depth k. We proceed with the induction for $\mathcal{O}(y)$. This shows that $\mathcal{O}$ is a convex combination of stabilizer operations of the form (3.27), where the measurements do not act on the k ancilla qudits.

Next, let us assume that the measurements in $\mathcal{O}$ act on the input system only. Then, using Lemma 3.4.4, we show that it is a convex combination of stabilizer operations where the measurements are given by mutually orthogonal stabilizer code projectors. To this end, we consider consecutive measurements along a branch and argue again via induction over the depth k of the tree. Assume that up to depth $k-1$, all consecutive measurements are mutually commuting. Let $X_{k-1} \subset$

$\mathbb{F}_d$ be the set of outcomes $x_{[k-1]}$ such that the Pauli measurement given by $P(x_k|x_{[k-1]})$ is not commuting with a previous measurement, say $P(x_t|x_{[t-1]})$ for $t < k$. Since by assumption the previous measurements mutually commute, we can write using Lemma 3.4.4

$$\begin{aligned} P(x_k \,|\, x_{[k-1]})P(x_{k-1} \,|\, x_{[k-2]}) \cdots P(x_t|x_{[t-1]}) \cdots P(x_1) \\ &= P(x_k \,|\, x_{[k-1]})P(x_t|x_{[t-1]})P(x_{k-1} \,|\, x_{[k-2]}) \cdots P(x_1) \\ &= d^{-1/2}V(x_{[k]})P(x_t|x_{[t-1]})P(x_{k-1} \,|\, x_{[k-2]}) \cdots P(x_1) \\ &= d^{-1/2}V(x_{[k]})P(x_{k-1} \,|\, x_{[k-2]}) \cdots P(x_t|x_{[t-1]}) \cdots P(x_1), \end{aligned} \tag{3.30}$$

for suitable Clifford unitaries $V(x_{[k]})$. Note that the remaining projectors are mutually commuting by assumption. As before, this implies that for all branches starting with $x_{[k-1]} \in X_{k-1}$, we can treat $x_k \equiv y$ as the outcome of a classical, uniformly distributed random variable Y and obtain new stabilizer operations $\mathcal{O}(y)$ by conditioning on its outcomes $y \in \mathbb{F}_d$:

$$K'(x_l, \ldots, x_{k+1}, y, x_{[k-1]}) := d^{1/2}K(x_l, \ldots, x_{k+1}, y, x_{[k-1]}), \quad x_{[k-1]} \in X_{k-1}, \tag{3.31}$$

$$K'(x_l, \ldots, x_k, x_{[k-1]}) := K(x_l, \ldots, x_k, x_{[k-1]}), \qquad x_{[k-1]} \notin X_{k-1}. \tag{3.32}$$

As in the previous argument, we have $\mathcal{O} = d^{-1} \sum_y \mathcal{O}(y)$ and proceeding with the induction for all $\mathcal{O}(y)$ shows that $\mathcal{O}$ can be written as a convex combination of stabilizer operations involving only mutually commuting, consecutive measurements.

Combining the above arguments shows that the initial stabilizer operation $\mathcal{O}$ defined in Eq. (3.27) is a convex combination of stabilizer operations $\mathcal{O}'$ where all consecutive measurements are mutually commuting and not acting on ancilla qudits. This implies that the mutually commuting projectors in every branch i of the SO $\mathcal{O}'$ define a stabilizer code projector P_i and trace-preservation requires that $\sum_i P_i = \mathbb{1}$. Taking the trace inner product with some P_j shows that

this can only be fulfilled if the projectors are mutually orthogonal. Hence, the terms in the convex combination are of the required type (3.25). The additional use of classical randomness simply allows for arbitrary convex combinations of stabilizer operations of type (3.25). □

3.4.2 Separation of CSP and SO in higher dimensions

To prove our main Theorem 6.3.1, we proceed by generalizing the results of Sec. 3.3 on almost-diagonal channels and derive constraints on almost-diagonal stabilizer operations using the normal form in Thm. 3.4.2.

As in Sec. 3.3, we define the convex set of *almost-diagonal channels* $\mathrm{AD}_n(d)$ as the set of quantum channels $\mathcal{E} : L((\mathbb{C}^d)^{\otimes n}) \to L((\mathbb{C}^d)^{\otimes n})$ that act on the computational basis in the following way:

$$\mathcal{E}(|0\rangle\langle 0|) = |+\rangle\langle +|\,, \qquad \mathcal{E}(|x\rangle\langle x|) = |x\rangle\langle x| \quad x \in \mathbb{F}_d^n \setminus 0, \tag{3.33}$$

where we denote $|+\rangle := d^{-n/2} \sum_{x\in\mathbb{F}_d^n} |x\rangle$. A high-level reason why $\mathrm{AD}_n(d)$ might be relevant for the separation of $\mathrm{CSP}_n(d)$ and $\mathrm{SO}_n(d)$ is given by the observation that $\mathrm{AD}_n(d)$ defines a *face* of the convex set of quantum channels. In particular, $\mathrm{AD}_n(d) \cap \mathrm{CSP}_n(d)$ is a face of the CSP polytope and thus lies in its boundary. To see this, consider the linear functional on quantum channels,

$$L(\mathcal{E}) := \frac{1}{d^n}\left(\langle +|\, \mathcal{E}(|0\rangle\langle 0|)\, |+\rangle + \sum_{x\neq 0} \langle x|\, \mathcal{E}(|x\rangle\langle x|)\, |x\rangle \right) \leq \frac{1}{d^n}(1 + (d^n - 1)\cdot 1) = 1, \tag{3.34}$$

with equality if and only if $\mathcal{E}$ satisfies Eq. (3.33). This shows that $\mathrm{AD}_n(d)$ is the intersection of a supporting hyperplane with the set of quantum channels, in particular it is a face.

As in the case $d = n = 2$, the subpolytope $\mathrm{AD}_n(d) \cap \mathrm{CSP}_n(d)$ of $\mathrm{CSP}_n(d)$ is isomorphic to a subpolytope $\mathrm{P}_n(d)$ of the n-qudit stabilizer polytope which we define in the following.

Definition 3.4.6. *Let $\mathrm{P}_n(d)$ be the polytope of matrices σ such that (1) σ is a convex combination of n-qudit stabilizer states orthogonal to $|0\rangle$, and (2) the diagonal entries are $\sigma_{x,x} = (d^n - 1)^{-1}\delta_{x\neq 0}$.*

Lemma 3.4.7. *Let $\mathcal{E} \in \mathrm{AD}_n(d)$ be an almost-diagonal quantum channel on n qudits. Then $\mathcal{E}$ is CSP if and only if it is of the form*

$$\mathcal{E}(\rho) = (d^n - 1)\, \sigma \circ \rho + \langle 0|\rho|0\rangle\, |+\rangle\langle +|, \tag{3.35}$$

where $\circ$ denotes the Hadamard *(or element-wise) product of two matrices and $\sigma \in \mathrm{P}_n(d)$. In particular, the polytopes $\mathrm{AD}_n(d)$ and $\mathrm{P}_n(d)$ are isomorphic.*

The proof of Lemma 3.4.7 is analogous to the case $d = n = 2$, i.e. Lemma 3.3.3, and is thus omitted. Lemma 3.4.7 implies that any CSP map fulfilling the constraints (3.33) is indeed "almost-diagonal" in the computational basis in the sense that $\mathcal{E}(|x\rangle\langle y|) \propto |x\rangle\langle y|$ except for $x = y = 0$. Hence, the matrix representation of $\mathcal{E}$ is a diagonal matrix with the first column (corresponding to $x = y = 0$) replaced by $(d^{-n}, \ldots, d^{-n})^\top$.

Lemma 3.4.8. *Any stabilizer operation in the subpolytope $\mathrm{AD}_n(d) \cap \mathrm{CSP}_n(d)$ is in the convex hull of stabilizer operations $\mathcal{O}$ defined through Lemma 3.4.7 by mixed stabilizer states*

$$\sigma = \frac{1}{d^n - 1} \sum_{K \in \mathcal{K}} |K|\, |s_K\rangle\langle s_K| \in \mathrm{P}_n, \tag{3.36}$$

where $\mathcal{K}$ is a disjoint partition of $\mathbb{F}_d^n \setminus 0$ by affine spaces $K \subset \mathbb{F}_d^n$ and $|s_K\rangle$ are stabilizer states supported on K.

Remark 3.4.9. *Note that not every σ of the form* (3.36) *gives rise to a stabilizer operation $\mathcal{E} \in \mathrm{AD}_n(d) \cap \mathrm{SO}_n(d)$. For stabilizer operations, only particular partitions $\mathcal{K}$ are allowed. These partitions exhibit a certain* tree structure*, as explained in Proposition 3.5.1 and [CB09].*

Moreover, not all such σ are extremal within the polytope $\mathrm{P}_n(d)$. For example, if a stabilizer operation contains the measurement of a Pauli operator and the measurement is not *followed by an operation that is conditioned on at least one of the measurement outcomes, then such an operation cannot be extremal. In this case, the measurement can be replaced by a convex combination of Clifford unitaries. This is a consequence of Lemma 3.C.1 in App. 3.C.*

To prove Lemma 3.4.8, we make use of the following Lemma which allows us to discard ancillary qubits for stabilizer operations in $\mathrm{AD}_n(d)$.

Lemma 3.4.10. *Assume $U \in \mathrm{Cl}_{n+k}(d)$ acts as $U(|0^k\rangle \otimes |x\rangle) = c_x\, |s_x\rangle \otimes |x\rangle$ for $c_x \in \mathbb{C}$, some k-qudit stabilizer state $|s_x\rangle$ and x is taking values in a subset $K \subset \mathbb{F}_d^n$. Then, there exists a diagonal Clifford unitary $D \in \mathrm{Cl}_n(d)$ and a subspace $M \subset \mathbb{F}_d^n$ such that the following identity holds for all $x, y \in K$:*

$$\mathrm{Tr}_{1,\ldots,k}\left(U|0^k\rangle\langle 0^k| \otimes |x\rangle\langle y|\, U^\dagger\right) = D\left(\sum_{j=1}^{|M|} Q_j\, |x\rangle\langle y|\, Q_j\right) D^\dagger. \tag{3.37}$$

Here, Q_j are the mutually orthogonal projectors onto the joint eigenspaces of $Z(z)$ for $z \in M$.

As we do not use this formulation in the following, we leave it to the reader to verify that the right hand side of Eq. (3.37) can also be written as

$$\sum_{j=1}^{|M|} Q_j\, |x\rangle\langle y|\, Q_j = \frac{1}{|M|} \sum_{z\in M} Z(z)\, |x\rangle\langle y|\, Z(z)^\dagger, \qquad \forall x, y \in \mathbb{F}_d^n. \tag{3.38}$$

Proof. The stabilizers of the states $|s_x\rangle$ can only differ by a character and hence we can find a Clifford $V \in \mathrm{Cl}_k(d)$ on the first system such that $V\,|s_x\rangle = |f(x)\rangle$ with $f(x) \in \mathbb{F}_d^n$ for all $x \in K$. Moreover, we find using the cyclicity of the partial trace:

$$\mathrm{Tr}_{1,\ldots,k}\left(U|0^k\rangle\langle 0^k| \otimes |x\rangle\langle y|\, U^\dagger\right) = \mathrm{Tr}_{1,\ldots,k}\left((V\otimes \mathbb{1})U|0^k\rangle\langle 0^k| \otimes |x\rangle\langle y|\, U^\dagger(V^\dagger \otimes \mathbb{1})\right). \tag{3.39}$$

Hence, we may without loss of generality assume that $U(|0^k\rangle \otimes |x\rangle) = c_x\, |f(x)\rangle \otimes |x\rangle$ for a suitable function f on $K \subset \mathbb{F}_d^k$. It is well-known that the Clifford subgroup which normalises the group of Pauli Z operators is given as the semi-direct product of diagonal Clifford unitaries and CX circuits (this follows for instance from the properties of the associated "Siegel parabolic subgroup" of the symplectic group $\mathrm{Sp}_{2n}(\mathbb{F}_2)$, see e.g. Ref. [Hei21]). Thus, the only Clifford unitaries which map computational basis states to computational basis states up to

phases are given by this normalizer and X gates. Since the second system is fixed for all $x \in K$, we can assume that the X gates act on the first system only and can thus be discarded using the cyclicity of the partial trace, cp. Eq. (3.39). Then, the form of diagonal Cliffords (see e. g. Ref. [DDM03]) implies that $c_x = \langle x| D |x\rangle$ for some diagonal Clifford unitary $D \in \mathrm{Cl}_n(d)$. Moreover, we can find a linear map $F \in \mathrm{GL}_{n+k}(\mathbb{F}_d)$ such that $F(0,x) = (f(x), x)$ and hence, f is linear.

Next, we argue that we can infer whether $\langle f(y)|f(x)\rangle$ is zero or one by a suitable measurement on the second system. To this end, note that this overlap is one exactly if $f(x) = f(y)$, i.e. $x - y \in \ker f$. This in turn the case if and only if $z \cdot (x-y) = 0$ for all $z \in M := (\ker f)^\perp$, hence if and only if$|x\rangle$ and $|y\rangle$ lie in the same joint eigenspace of the stabilizer group $\{Z(z) \mid z \in M\}$. Note that any computational basis state always lies in one of the eigenspaces. Let Q_j for $j = 1, \ldots, |M|$ be the projectors on these stabilizer codes. We thus find

$$D\left(\sum_{j=1}^{|M|} Q_j \,|x\rangle\langle y|\, Q_j\right) D^\dagger = \bar{c}_y c_x \,\langle f(y)|f(x)\rangle\, |x\rangle\langle y| = \mathrm{Tr}_{1,\ldots,k}\left(U|0^k\rangle\langle 0^k| \otimes |x\rangle\langle y|\, U^\dagger\right). \tag{3.40}$$

□

Proof of Lemma 3.4.8. Assume that $\mathcal{O} \in \mathrm{AD}_n(d)$ is a stabilizer operation. Without loss of generality, we can assume that $\mathcal{O}$ is extremal in $\mathrm{SO}_n(d)$, since any $\mathcal{O} \in \mathrm{AD}_n(d) \cap \mathrm{SO}_n(d)$ can be written as a convex combination of extremal SO in $\mathrm{AD}_n(d)$ by the same argument as in Lemma 3.3.4. By Proposition 3.4.2, we can thus assume that $\mathcal{O}$ has the following form

$$\mathcal{O}(\rho) = \mathrm{Tr}_{1,\ldots,k} \sum_{i=1}^{N} U_i \left(|0^k\rangle\langle 0^k| \otimes P_i \rho P_i\right) U_i^\dagger, \tag{3.41}$$

where the P_i are mutually orthogonal stabilizer code projectors of rank d^{n-r_i} on the input system, and the U_i are Clifford unitaries conditioned on the measurement outcomes. Let $\tilde{\mathcal{O}} \in \mathrm{SO}_{n,n+k}$ be the SO given by Eq. (3.41) without the partial trace. Since the defining condition (3.14) for AD_n requires that the reduced state $\mathrm{Tr}_{1,\ldots,k}\,\tilde{\mathcal{O}}(|x\rangle\langle x|)$ is pure

for all $x \in \mathbb{F}_d^n$, it is necessary that $\tilde{\mathcal{O}}(|x\rangle\langle x|) = \rho_x \otimes |x\rangle\langle x|$ for $x \neq 0$ and $\tilde{\mathcal{O}}(|0\rangle\langle 0|) = \rho_0 \otimes |+\rangle\langle +|$ else. Similar to Eq. (3.19) before, this requires that

$$U_i \left(|0^k\rangle \otimes P_i |0^n\rangle \right) \propto |s_{i,0}\rangle \otimes |+\rangle \,, \tag{3.42}$$

$$U_i \left(|0^k\rangle \otimes P_i \, |x\rangle \right) \propto |s_{i,x}\rangle \otimes |x\rangle \,, \quad x \neq 0, \tag{3.43}$$

where proportionality can also mean that the RHS vanishes.

Let i be such that Eq. (3.42) holds with non-vanishing constant, without loss of generality $i = 1$. If any of the P_i were non-diagonal, a standard argument (cp. Lem. 3.A.3 in App. 3.A) would show that there exist distinct computational basis states $|x\rangle \neq |y\rangle$ such that $P_i \, |x\rangle = P_i \, |y\rangle \neq 0$, which would contradict Eqs. (3.42) and (3.43). Thus, all P_i are diagonal. Moreover, if P_1 had rank larger than 1, there would be a $x \neq 0$ such that $P_1 \, |x\rangle = |x\rangle$ is orthogonal to $P_1 \, |0\rangle = |0\rangle$. But then, the second factor of $U_1(|0\rangle \otimes P_1 \, |x\rangle)$ has to be an X eigenstate, in contradiction to Eq. (3.43). Hence, the projectors have the form

$$P_1 = |0\rangle\langle 0| \,, \qquad P_i = \sum_{x \in K_i} |x\rangle\langle x| \,, \tag{3.44}$$

where $0 \notin K_i \subset \mathbb{F}_d^n$ is an affine subspace not containing zero. Then, orthogonality of the P_i implies that the K_i form a disjoint partition of $\mathbb{F}_d^n \setminus 0$.

Note that we can assume that $U_1 = V \otimes H^{\otimes n}$ (up to Z operators). Therefore, we can simply trace out the ancilla for the first term. For $i > 1$, consider the conditional Clifford unitary U_i which acts on the code space K_i as $U_i \, |0\rangle \otimes |x\rangle = c_i(x) \, |s_{i,x}\rangle \otimes |x\rangle$ where $c_i(x) \in \mathbb{C}$. Then, we can use Lemma 3.4.10 to replace this action by a diagonal Clifford D_i and m_i mutually orthogonal diagonal stabilizer code projectors on the *input system*. We can write these projectors as

$$Q_j^i = \sum_{x \in A_j^i} |x\rangle\langle x| \,, \qquad j = 1, \ldots, m_i, \tag{3.45}$$

where $A_j^i \subset \mathbb{F}_d^n$ are suitable affine subspaces (which might contain zero), forming a disjoint partition of $\mathbb{F}_d^n$. Now consider

$$P_{i,j} := Q_j^i P_i = \sum_{x \in A_j^i \cap K_i} |x\rangle\langle x| \,. \tag{3.46}$$

Here, $K_j^i := A_j^i \cap K_i$ is an affine subspace not containing zero. Note that $\{K_j^i\}_{j=1,\dots,m_i}$ is a disjoint partition of K_i and thus, the affine subspaces $\mathcal{K} := \{K_j^i \mid i = 2,\dots,N,\ j = 1,\dots,m_i\}$ obtained in this way form a disjoint partition of $\mathbb{F}_d^n \setminus 0$. Finally, we can write

$$\mathcal{O}(\rho) = |+\rangle\langle 0|\,\rho\,|0\rangle\langle +| + \sum_{i=2}^{N}\sum_{j=1}^{m_i} D_i Q_j^i P_i \rho P_i Q_j^i D_i^\dagger = |+\rangle\langle 0|\,\rho\,|0\rangle\langle +| + (d^n-1)\,\sigma\circ\rho, \tag{3.47}$$

where

$$\sigma = \frac{1}{d^n-1}\sum_{(i,j)} |K_j^i|\ |s_{i,j}\rangle\langle s_{i,j}|\,, \qquad |s_{i,j}\rangle := |K_j^i|^{-\frac{1}{2}} \sum_{x\in K_j^i} D_i\,|x\rangle\,. \tag{3.48}$$

To see that this is indeed an AD_n channel, note that any of the $|s_{i,j}\rangle$ is a stabilizer state and as the K_j^i form a disjoint partition of $\mathbb{F}_d^n \setminus 0$, σ is a proper convex combination and hence in $\mathrm{SP}_n(d)$. Finally, we have $\langle 0|\,\sigma\,|x\rangle = 0$ for all $x \in \mathbb{F}_d^n$ and for $x \neq 0$:

$$(d^n-1)\,\langle x|\,\sigma\,|x\rangle = \sum_{(i,j)} |K_j^i|\,|\,\langle x|s_{i,j}\rangle\,|^2 = \sum_{(i,j)} \mathbf{1}_{K_j^i}(x). \tag{3.49}$$

Since any $x \neq 0$ is in exactly one affine subspace K_j^i, we thus find $\sigma \in \mathrm{P}_n(d)$. □

In Section 3.3, Lemma 3.4.11 has been proven for the case $n = d = 2$. Here, we treat the general case.

Lemma 3.4.11. *The matrix λ with elements*

$$\lambda_{x,tx} = \lambda_{tx,x} = (d^n-1)^{-1}\delta_{t,1}, \quad \forall x \in \mathbb{F}_d^n, t \in \mathbb{F}_d, \ \ \lambda_{x,y} = d^{-1}(d^n-1)^{-1}, \ \ \forall\, 0 \neq x \neq y$$

is the unique maximizer in P_n of the linear function $L : \sigma \mapsto \langle +|\,\sigma\,|+\rangle$ with $L(\lambda) = 1/d$. In particular, λ is a vertex of P_n.

Proof of Lemma 3.4.11. For any $\sigma \in \mathrm{P}_n$, we have

$$L(\sigma) = \langle +|\sigma|+\rangle = \sum_s p_s |\langle +|s\rangle|^2, \tag{3.50}$$

where $|s\rangle$ ranges over stabilizer states orthogonal to $|0\rangle$. Among those, the inner product $|\,\langle +|s\rangle\,|^2$ is maximal and equal to $1/d$ exactly for

states such that $\langle x|s\rangle$ is proportional to an indicator function on an affine space K of codimension 1 with $0 \neq K$. We call K a *proper affine hyperplane*.

We can write any affine hyperplane as $K = V + w$ where V is a linear subspace of codimension 1 and w is an affine shift. Those are only determined modulo V, hence there are $|\mathbb{F}_d^n/V| = d$ many. The condition $0 \neq K$ implies that the shift cannot be trivial, eliminating one possibility. The number of linear subspaces of codimension 1 is given by the Gaussian binomial coefficient $\binom{n}{n-1}_d = \frac{1-d^n}{1-d}$, hence the number of proper affine hyperplanes is $(d-1)\frac{1-d^n}{1-d} = d^n - 1$.

Define λ to be the uniform convex combination of all maximizing stabilizer states $|K\rangle$ given by indicator functions on the proper affine hyperplanes K. For any $x \in \mathbb{F}_d^n \setminus 0$, the diagonal entry $\lambda_{x,x}$ is the overlap $\langle x|K\rangle \propto \mathbf{1}_K(x)$ averaged over K. As $\mathrm{GL}(\mathbb{F}_d^n)$ acts transitively on both the non-zero points in $\mathbb{F}_d^n$ and the affine spaces not containing zero, $\lambda_{x,x}$ cannot depend on $x \neq 0$. Since $\operatorname{Tr}\lambda = 1$ and $\lambda_{0,0} = 0$, we thus find $\lambda_{x,x} = (d^n - 1)^{-1}$. For the off-diagonal entries $\lambda_{x,y}$ with $x \neq y$, we can argue similarly. First, as no K contains zero, we have $\lambda_{x,0} = \lambda_{0,x} = 0$. Moreover, if $x \in K$, no non-trivial multiple of x is in K, thus $\lambda_{x,tx} = \lambda_{tx,x} = 0$ for $t \neq 1$. In any other case, $\{x, y\}$ is linearly independent and transitivity again implies that $\lambda_{x,y}$ cannot depend on (x, y). There are in total $(d^n - 1)(d^n - d)$ many of these pairs. By construction, $L(\lambda) = 1/d$, and writing out this condition then yields $\lambda_{x,y} = d^{-1}(d^n - 1)^{-1}$.

It remains to be shown that this solution is unique. To this end, we claim that the $d^n - 1$ indicator functions $\mathbf{1}_K$ on the proper affine hyperplanes K are linearly independent. As the main diagonal of the density matrix of any stabilizer state $|K\rangle$ is proportional to $\mathbf{1}_K$, the $d^n - 1$ constraints $\sigma_{x,x} = (d^n - 1)^{-1}$ for $x \neq 0$ defining P_n then single out the above constructed λ.

To prove the claim, we apply the standard (cyclic) Fourier transform on $\mathbb{F}_d^n$. Clearly, the set of indicator functions on proper affine hyperplanes $\{\mathbf{1}_K\}$ is linearly independent if and only if their Fourier transforms are. The image of the indicator function on $K = V + w$ is

a function with support on $V^\perp$ and values proportional to the additive character

$$\chi : V^\perp \to \mathbb{C}, \qquad x \mapsto \omega^{w\cdot x}.$$

As we assume w to be non-trivial, varying w results in the set of non-trivial characters on the one-dimensional subspace $V^\perp$. Thus, the set $\{\mathbf{1}_K\}$ maps to the set of non-trivial characters on the one-dimensional subspaces of $\mathbb{F}_d^n$. On a fixed subspace $V^\perp$, the non-trivial characters are linearly independent and this is still true for their restriction to the non-zero points $V^\perp \setminus 0$. Since the non-zero points of one-dimensional subspaces form a disjoint partition of $\mathbb{F}_d^n \setminus 0$, the set of all non-trivial characters of one-dimensional subspaces is also linearly independent. □

From the proof and Lemma 3.4.7 it is clear that the matrix λ defines a CSP channel which should be understood as a generalisation of the Λ channel given for $n = d = 2$ in Sec. 3.3. For qubits, $d = 2$, this channel reads as follows:

$$\Lambda(\rho) := \rho_{00} \left|+\right\rangle\!\left\langle+\right| + \sum_{x\in\mathbb{F}_2^n\setminus 0} \rho_{xx} \left|x\right\rangle\!\left\langle x\right| + \frac{1}{2} \sum_{\substack{x,y\in\mathbb{F}_2^n\setminus 0 \\ x\neq y}} \rho_{xy} \left|x\right\rangle\!\left\langle y\right|, \qquad \rho_{xy} := \left\langle x\right| \rho \left|y\right\rangle. \tag{3.51}$$

The $n \geq 2$ case in our main Theorem 6.3.1 now follows from the straightforward observation that the linear functional L is always strictly less than its maximum $1/d$ on elements of the form given in Lemma 3.4.8, in particular on stabilizer operations.

Corollary 3.4.12 ($\mathrm{SO}_n \cap \mathrm{AD}_n \neq \mathrm{CSP}_n \cap \mathrm{AD}_n$)**.** *For $n \geq 2$, the value of the linear functional L on $\mathrm{SO}_n \cap \mathrm{AD}_n$ is strictly smaller than $1/d = L(\lambda)$. In particular, $\mathrm{SO}_n \cap \mathrm{AD}_n \neq \mathrm{CSP}_n \cap \mathrm{AD}_n$.*

Proof. Consider a stabilizer operation with $\sigma \in P_n$ as constructed in Lemma 3.4.8, i.e.

$$\sigma = \frac{1}{d^n - 1} \sum_{K\in\mathcal{K}} |K| \left|s_K\right\rangle\!\left\langle s_K\right|. \tag{3.52}$$

Since $n \geq 2$, the disjoint partition $\mathcal{K}$ of $\mathbb{F}_d^n \setminus \{0\}$ cannot only contain affine subspaces K of codimension 1, so $|\langle +|s_K\rangle|^2 \leq 1/d$ and $|\langle +|s_K\rangle|^2 < 1/d$ for at least one $K \in \mathcal{K}$. Hence,

$$\langle +|\sigma|+\rangle = \frac{1}{d^n-1} \sum_{K\in\mathcal{K}} |K|\,|\langle +|s_K\rangle|^2 < \frac{1}{d^n-1} \sum_{K\in\mathcal{K}} |K| \frac{1}{d} = \frac{1}{d}. \quad (3.53)$$

□

Remark 3.4.13. *To get a quantitative statement about the separation of* SO_n *and* CSP_n *within the polytope* AD_n*, we give a upper bound for*

$$\max_{\sigma\in \mathrm{SO}_n \cap \mathrm{AD}_n} \langle +|\sigma|+\rangle. \quad (3.54)$$

As in Eq. 3.53, we have

$$\langle +|\sigma|+\rangle = \frac{1}{d^n-1} \sum_{K\in\mathcal{K}} |K|\,|\langle +|s_K\rangle|^2 \leq \frac{1}{d^n-1} \sum_{K\in\mathcal{K}} |K| \frac{|K|}{d^n} = \frac{1}{d^n(d^n-1)} \sum_{K\in\mathcal{K}} |K|^2. \quad (3.55)$$

The RHS gets large when affine subspaces in the disjoint partition $\mathcal{K}$ *of* $\mathbb{F}_2^n \setminus \{0\}$ *have large cardinality. However, the partition* $\mathcal{K}$ *must be chosen according to Thm. 3.4.2. Thus, all projectors* $|s_K\rangle\langle s_K|$ *belong to a projective measurement, which also contains the measurement of* $|0^n\rangle\langle 0^n|$*, due to the proof of Lemma 3.4.8. We conjecture that such a projective measurement which maximizes* (3.55) *has the following form:*

(i) Measure the first qudit in the computational basis.

(ii) If $x \neq 0$ *is measured, do nothing. If* 0 *is measured, measure the second qudit in the computational basis. Continue in this fashion until all qubits are measured.*

(iii) If 0 *has been measured on every qudit, apply a Hadamard gate to every qudit.*

Then, $|s_K\rangle = \sum_{x\in K} |x\rangle$ *and every* K *is of the form*

$$K = xe_i + (0^i \oplus \mathbb{F}_d^{n-i}) \qquad \text{with} \qquad x \in \{1,\ldots,d-1\}, \quad |K| = d^{n-i}, \quad (3.56)$$

where e_i is the i-th standard basis vector for $i = 0, \ldots, n-1$. Thus, we have

$$\sum_{K\in\mathcal{K}} |K|^2 = \sum_{k=0}^{n-1} (d-1)d^{2k} = (d-1)\frac{d^{2n-2}}{d^2-1}, \tag{3.57}$$

where we used that the expression is a geometric sum. Hence,

$$\begin{aligned}\langle +|\sigma|+\rangle &= \sum_{K\in\mathcal{K}} |K|\frac{|K|}{d^n} = \frac{1}{d^n(d^n-1)} \sum_{K\in\mathcal{K}} |K|^2 \\ &= \frac{(d-1)d^{2n-2}}{d^n(d^n-1)(d^2-1)} = \frac{(d-1)d^{n-2}}{(d^n-1)(d^2-1)} \approx \frac{1}{d^3}. \end{aligned} \tag{3.58}$$

As the above stabilizer operation gives a upper bound on $\max_{\sigma\in \mathrm{SO}_n \cap \mathrm{AD}_n} \langle +|\sigma|+\rangle$, this shows a separation between $\mathrm{SO}_n \cap \mathrm{AD}_n$ and $\mathrm{CSP}_n \cap \mathrm{AD}_n$ which depends, however, only on d and not on n.

3.4.3 Equality of SO and CSP in the single-qudit case

In this section, we will prove that CSP-channels coincide with stabilizer operations in the single-qudit case. More precisely, we will show that every extremal CSP-map is a Pauli measurement followed Clifford unitaries conditioned on the possible measurement outcomes. In the proof we will make use of the polar form of CSP-maps, Eq. (3.69), App. 3.5 (for more details, see [Hei21]).

To prove the statement, we will use the following auxiliary lemma:

Lemma 3.4.14. *Suppose $\mathcal{E} \in \mathrm{CSP}_n$ is a CSP map given in the polar form* (3.69)

$$\mathcal{E} = \sum_i c_i U_i P_i \cdot P_i U_i^\dagger, \qquad \text{where } c_i > 0 \text{ for all } i. \tag{3.59}$$

Assume that there exists an index pair (k, ℓ) with $P_k = P_\ell$ but $U_k P_k \neq U_\ell P_\ell$. Then, $\mathcal{E}$ is not extremal.

Proof. Since $\mathcal{E} \in \mathrm{CSP}_n$, the projectors P_i satisfy the TP-condition (3.70)

$$\mathbb{1} = c_k P_k + c_\ell P_\ell + \sum_{k\neq i\neq \ell} c_i P_i \tag{3.60}$$

and therefore

$$\mathbb{1} = (c_k + c_\ell)P_k + \sum_{k\neq i\neq \ell} c_i P_i \quad \text{and} \quad \mathbb{1} = (c_k + c_\ell)P_\ell + \sum_{k\neq i\neq \ell} c_i P_i. \tag{3.61}$$

Hence, $\mathcal{E}$ is a convex combination $\mathcal{E} = \frac{c_k}{c_k+c_\ell}\mathcal{E}_k + \frac{c_\ell}{c_k+c_\ell}\mathcal{E}_\ell$ of the two distinct CSP-channels

$$\begin{aligned} \mathcal{E}_k &= (c_k + c_\ell)U_k P_k \cdot P_k U_k + \sum_{k\neq i\neq \ell} c_i U_i P_i \cdot P_i U_i^\dagger, \\ \mathcal{E}_\ell &= (c_k + c_\ell)U_\ell P_\ell \cdot P_\ell U_\ell + \sum_{k\neq i\neq \ell} c_i U_i P_i \cdot P_i U_i^\dagger, \end{aligned} \tag{3.62}$$

so $\mathcal{E}$ cannot be extremal. □

Theorem 3.4.15. *Let $\mathcal{E} \in \mathrm{CSP}_1$ be an extremal CSP map on a single qudit of prime dimension d. Then, either $\mathcal{E} = U \cdot U^\dagger$ for some Clifford unitary U or $\mathcal{E}$ is of the form*

$$\mathcal{E} = \sum_{i=1}^{d} U_i P_i \cdot P_i U_i^\dagger, \tag{3.63}$$

where $\{P_i\}$ are the d mutually orthogonal stabilizer code projectors associated to the eigenspaces of a Pauli operator and $\{U_i\}$ are Clifford unitaries. Since such a channel $\mathcal{E}$ can be realised via stabilizer operations, it follows $\mathrm{SO}_1 = \mathrm{CSP}_1$.

Proof. Using the characterization of completely stabilizer preserving maps, cf. Eq. (3.69), we may assume that a 1-qudit CSP channel is of the form

$$\mathcal{E} = d\sum_i \lambda_i U_i P_i \cdot P_i U_i^\dagger + \sum_j \hat{\lambda}_j V_j \cdot V_j^\dagger \tag{3.64}$$

for coefficients $\lambda_i, \hat{\lambda}_j \geq 0$ with $\sum_i \lambda_i + \sum_j \hat{\lambda}_j = 1$, Clifford unitaries U_i, V_j and stabilizer code projectors P_i which satisfy the TP-condition (3.70) :

$$\mathbb{1} = \mathcal{E}^\dagger(\mathbb{1}) = d\sum_i \lambda_i P_i. \tag{3.65}$$

Since any channel that simply conjugates the input with a Clifford unitary U is already an extremal CSP channel, $\mathcal{E}$ can only be extremal if (1) there is exactly one non-zero $\hat{\lambda}_j$ and $\lambda_i = 0$ for all i (which means that $\mathcal{E} = U \cdot U^\dagger$ for some Clifford unitary U), or (2) $\hat{\lambda}_j = 0$ for all j.

In the second case, note that for $n = 1$, all stabilizer projectors have rank 1 and project onto a stabilizer state. There are in total $d(d+1)$ stabilizer states $|\phi_{i,a}\rangle$ which form a complete set of mutually unbiased bases, in particular for any $a = 1, \ldots, d+1$ the set $\{|\phi_{i,a}\rangle\}_i$ is an orthonormal basis. By Lemma 3.4.14, we can assume that every projector $|\phi_{i,a}\rangle\langle\phi_{i,a}|$ only occurs at most once in $\mathcal{E}$. Thus, grouping the projectors by their basis, we can write the CSP channel $\mathcal{E}$ as

$$\mathcal{E} = d \sum_{a=1}^{d+1} \sum_{i=1}^{d} \lambda_{i,a} U_{i,a} \, |\phi_{i,a}\rangle\langle\phi_{i,a}| \cdot |\phi_{i,a}\rangle\langle\phi_{i,a}| \, U_{i,a}^\dagger. \tag{3.66}$$

Since every basis $\{|\phi_{i,a}\rangle\}_i$ is the eigenbasis of a (non-trivial) Pauli operator w and further $\mathrm{Tr}(w \, |\phi_{i,a'}\rangle\langle\phi_{i,a'}|) = 0$ for $a \neq a'$, taking the trace inner product of Eq. (3.65) multiplied with w implies

$$0 = \sum_{i=1}^{d} \lambda_{i,a} \omega^i, \tag{3.67}$$

which forces the $\lambda_{i,a}$ to be either identically zero or independent of i. Setting $\tilde{\lambda}_a = d^{-1} \lambda_{i,a}$, we thus arrive at a convex combination of $\mathcal{E}$ into CSP channels $\mathcal{E}_a$:

$$\mathcal{E} = \sum_{a=1}^{d+1} \tilde{\lambda}_a \mathcal{E}_a, \qquad \mathcal{E}_a := \sum_{i=1}^{d} U_{i,a} \, |\phi_{i,a}\rangle\langle\phi_{i,a}| \cdot |\phi_{i,a}\rangle\langle\phi_{i,a}| \, U_{i,a}^\dagger. \tag{3.68}$$

Hence, extremality of $\mathcal{E}$ implies that it is of the desired form. □

3.5 Additional properties of CSP channels and examples

In this section, we derive additional properties of completely stabilizer-preserving channels which are not directly used to show the main result of this paper. We characterise CSP channels in terms of certain generalised stabilizer measurements and adaptive Clifford unitaries. This is what we call the *polar form* and has been used in Sec. 3.4.3

as well as in the simulation protocol of [SRP+21]. We then use this characterization to compile a list of examples of CSP channels.

By Lemma 3.2.5, completely stabilizer-preserving maps are in bijection with the subset of the bipartite $2n$-qudit stabilizer polytope fulfilling the TP condition. Notably, bipartite stabilizer states have a special structure that can be exploited to bring them into a standard form which we call the *polar form*. It is given by $|s\rangle = d^{k/2} UP \otimes \mathbb{1} \big|\phi^+\big\rangle$ for a Clifford unitary $U \in \mathrm{Cl}_n(d)$ and a stabilizer code projector P of rank d^{n-k}. Note that from this form, one can immediately derive the Schmidt rank of $|s\rangle$ as $\log_d \operatorname{rank}(P) = n - k$. While this fact seems to be folk knowledge in the relevant community and related results can be found in Refs. [How73, How88, FCY+04], we have been unable to find an explicit formulation in the literature. A proof of this fact can be found in the PhD thesis of one of the authors [Hei21, Sec. 12.3.2].

Proposition 3.5.1. *The $2n$-qudit state $|s\rangle \in (\mathbb{C}^d)^{\otimes 2n}$ is a stabilizer state if and only if there is a Clifford unitary $U \in \mathrm{Cl}_n(d)$ and a stabilizer code $P \in \mathrm{STAB}(k, n)$ such that $|s\rangle = d^{k/2} UP \otimes \mathbb{1} \big|\phi^+\big\rangle$.*

While the projective part in the polar form of a stabilizer state is unique, the unitary part is not. This is because replacing the Clifford unitary by $U \mapsto UV$ where V acts trivially on the code space gives an equivalent presentation of the state. Technically, this means that the unitary part is unique *up to the left Clifford stabilizer* of the stabilizer code.

Using the polar form, the polytope of CSP maps can be characterised as follows: The $\mathrm{SP}_{2n}(d)$ polytope corresponds under the inverse Choi-Jamiołkowski isomorphism to the polytope which is spanned by channels with a single stabilizer Kraus operator $d^{k/2} UP$. Hence, any CSP map is of the form

$$\mathcal{E} = \sum_{i=1}^{r} \lambda_i \frac{d^n}{\operatorname{rank} P_i} U_i P_i \cdot P_i U_i^\dagger, \tag{3.69}$$

where the λ_i form a probability distribution. However, Eq. (3.69) only defines a valid CSP map $\mathcal{E}$ if it is trace-preserving. We can cast the TP condition into an appealing form: $\mathcal{E}$ is a CSP map if and only if

in addition to Eq. (3.69), it fulfils

$$\mathbb{1} = \mathcal{E}^\dagger(\mathbb{1}) = \sum_{i=1}^{r} \frac{d^n \lambda_i}{\operatorname{rank} P_i} P_i. \tag{3.70}$$

Thus, a sufficient and necessary condition for a convex combination of stabilizer Kraus operators to define a CSP map is that the rescaled projective parts $\tilde{P}_i := (d^n \lambda_i / \operatorname{rank} P_i) P_i$ form a POVM. In this context, the CSP channel $\mathcal{E}$ in Eq. (3.69) can be seen as the quantum instrument associated with the stabilizer POVM $\{\tilde{P}_i\}$ combined with the application of Clifford unitaries U_i conditioned on outcome i.

A possible solution to Eq. (3.70) is a *syndrome measurement*, i.e. the POVM that is defined by the measurement of a set of mutually commuting Pauli operators (cp. example 3 below). Then, the corresponding CSP channel $\mathcal{E}$ is a stabilizer operation. However, as stabilizer operations are also allowed to use auxiliary qubits, they can effectively induce more complicated POVMs that fulfil Eq. (3.70). A priori, it is thus not clear whether CSP channels are different from stabilizer operations (this is, of course, answered by our main theorem 6.3.1). Interestingly, it even seems to be difficult to find solutions to Eq. (3.70) in terms of admissible stabilizer codes P_i and coefficients λ_i. In particular, one could think of arranging overlapping codes with the right weights in non-trivial ways such that they yield the identity on Hilbert space. Indeed, an example of a CSP channel defined via overlapping stabilizer codes is the Λ channel used for our main argument, see also App. 3.B. Note that given a set of stabilizer codes, it is in principle possible to decide whether there exist coefficients such that Eq. (3.70) holds by solving a linear system of equations which depends on the structure of code overlaps.

Finally, let us give some examples of CSP maps:

1. *Mixed Clifford channels.* Take $P_i \equiv \mathbb{1}$, then $d^n / \operatorname{rank} P_i = 1$ and Eq. (3.70) is trivially fulfilled for any convex combination.

2. *Dephasing in a stabilizer basis.* Take a basis of stabilizer states, and let P_i be the rank-one projectors onto the basis. A uniform convex combination $\lambda_i = d^{-n}$ of these fulfils the TP condition

Eq. (3.70). Such a channel corresponds to a dephasing in the chosen basis, followed by the potential application of conditional Clifford unitaries U_i depending on the basis measurement outcome i.

3. *Dephasing in stabilizer codes.* More generally, take an arbitrary stabilizer group $S = \langle g_1, \ldots, g_k \rangle$ and let P_i be all d^k orthogonal stabilizer codes corresponding to different phases of the generators and $\lambda_i = d^{-k}$. This defines a POVM ("syndrome measurement").

4. *Reset channels.* Let $s \in \mathrm{STAB}(n)$ be an arbitrary stabilizer state and consider the channel which replaces every input by s, i.e. $\mathcal{R}_s : X \mapsto \mathrm{Tr}(X)s$. It is clearly CSP and is a special cases of the second example where $|s\rangle$ is completed to a stabilizer basis and the Clifford unitaries are chosen such that all basis elements are mapped to $|s\rangle$.

3.6 Summary and open questions

In this work, we have studied and compared two classes of free operations in the resource theory of magic state quantum computing, namely completely stabilizer-preserving (CSP) channels and stabilizer operations (SO). Our main result shows that the set of multi-qudit CSP channels is always strictly larger than its subset of stabilizer operations. In the single-qudit case, however, the two classes coincide. Thus, our result is in analogy with the well-known fact from entanglement theory that LOCC operations are contained but not equal to the set of separable quantum channels.

Our proof strategy is simplified by the observation that it is sufficient to show the separation of CSP and SO in a suitable subspace. Having derived restrictions on the form of CSP and SO channels in this subspace, we then give a linear functional which is able to separate the two sets. In particular, we explicitly construct a CSP channel which is the unique maximizer of said functional and thus extremal in CSP.

As an auxiliary result, we restrict the form of Kraus operators of extremal stabilizer operations. In particular, this implies that stabilizer operations can be realised in a finite number of rounds. This is in contrast to entanglement theory, where the analogous LOCC operations become strictly more powerful with the number of rounds.

In our operational definition of SO, we intentionally allow for arbitrary classical control logic. As laid out in Sec. 3.2.2, this is implicit in the axiomatic definition of CSP and a separation would otherwise be trivial. However, as our proof does not depend on the details of the classical control, the separation still holds if we restrict the latter to efficient classical algorithms. For CSP, this has to be understood in the sense of Sec. 3.5, i.e. as efficient classical processing of the outcomes of generalised stabilizer POVMs and control of adaptive Clifford operations.

Some magic monotones, such as the *dyadic negativity* can be connected to classical simulation algorithms [SRP+21]. These allow to efficiently simulate a restricted class of CSP channels which is, however, strictly smaller than CSP with efficient classical control. The main reason for this is that it is not clear how to efficiently simulate the generalised stabilizer POVMs introduced in Sec. 3.5. Therefore, additional assumptions on these POVMs are necessary. However, it is plausible that CSP with these restricted POVMs is still strictly larger than SO. Hence, we expect that the algorithm by [SRP+21] allows for simulation beyond the Gottesman-Knill theorem. A thorough analysis of the simulability of CSP channels and comparison with the Gottesman-Knill theorem is left for future work.

Finally, we think that our result will stimulate further research in the resource theory of magic state quantum computing. The axiomatic approach to free operations has the advantage that it is possible to directly apply results from general resource theory and obtain explicit bounds on e.g. state conversion and distillation rates [VMGE14, Liu19, FL20, SRP+21, WWS20]. For the case of stabilizer-preserving channels, it is also known that the theory is asymptotically reversible

[LW22]. Here, it would be interesting to investigate which results still hold when the set of free operations is restricted to CSP. Moreover, if "free" shall have an operational meaning, then the question of simulability and the power of classical control will have to be discussed.

Our separation result opens the possibility that tasks like magic state distillation show a gap in the achievable rates between CSP channels and stabilizer operations. Again, this question is motivated from entanglement theory, where a significant separation between separable channels and LOCC operations for e.g. entanglement conversion is known [CCL12].

3.7 Acknowledgements

We would like to thank James Seddon and Earl Campbell for discussions which helped initializing this project. Furthermore, we thank Felipe Montealegre Mora for many discussions during the various stages of this work, and Mateus Araújo for helpful input on ancilla-assisted operations. This work has been supported by Germany's Excellence Strategy – Cluster of Excellence *Matter and Light for Quantum Computing (ML4Q)* EXC2004/1, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within the Emmy Noether program (grant number 441423094) and the Priority Program CoSIP, the German Federal Ministry for Education and Research through the Quantum Technologies program (QuBRA, QSolid) and the German Federal Ministry for Economic Affairs and Climate Action (ProvideQ).

3.A Miscellaneous facts on stabilizer states

Here, we state a fact (Proposition 3.A.2) on *stabilizer bases* that seems to be widely known, but for which we could not find a direct reference. It is used in the proof of Lemma 3.3.3.

Let S be a stabilizer group on n qudits of size $|S| = d^n$. There is a unique (up to phases) joint eigenbasis $\{|\alpha_i\rangle\}_i$ of all elements s of S.

Concretely: The eigenvalue equations

$$\chi_i(s)s|\alpha_i\rangle = |\alpha_i\rangle, \qquad s \in S \tag{3.71}$$

establish a one-one correspondence between the set of characters of S and elements of the common eigenvectors. Bases arizing this way are called *stabilizer bases*.

The argument uses basic notions from the description of Pauli operators and stabilizer states in terms of discrete symplectic vector spaces [Gro06]. In particular, two Pauli operators $w(a), w(b)$ for commute if and only if the *symplectic inner product*

$$[a, b] = \sum_{i=1}^{n} (a_z)_i (b_x)_i - \sum_{i=1}^{n} (a_x)_i (b_z)_i$$

is zero (as an element of $\mathbb{F}_d$). A subset $M \subset \mathbb{F}_d^{2n}$ is *isotropic* if the symplectic inner product vanishes between any two elements of M. An isotropic subspace M is *maximal* if $\dim M = n$. Witt's Lemma implies that every isotropic set is contained in a maximal isotropic subspace.

Lemma 3.A.1. *Let $M \subset \mathbb{F}_d^{2n}$ be an isotropic set. There is a stabilizer basis such that all pure states in the linear span of $\{w(a) \,|\, a \in M\}$ belong to that basis.*

Proof. Choose a basis $b_1, \ldots, b_n$ for some maximal isotropic subspace containing M. Then the operators $w(b_1), \ldots, w(b_n)$ generate a stabilizer group S of size d^n. Their unique common eigenbasis is a stabilizer basis. By construction, any pure state $|\psi\rangle\langle\psi|$ contained in the span of $\{w(a) \,|\, a \in M\}$ commutes with the elements in S and is thus a joint eigenvector of all elements in S, which implies that $|\psi\rangle$ belongs to the stabilizer basis of S. □

Proposition 3.A.2. *Let $|\psi\rangle \in (\mathbb{C}^d)^{\otimes n_1} \otimes (\mathbb{C}^d)^{\otimes n_2}$ be a bi-partite stabilizer state. Let $\{|\alpha_1\rangle, \ldots, |\alpha_{d^{n_1}}\rangle\}$ be a stabilizer basis on the first subsystem. Then there is a stabilizer basis on the second subsystem such that each of the partial contractions*

$$|\beta_i\rangle = (\langle\alpha_i| \otimes \mathbb{1})|\psi\rangle \in (\mathbb{C}^d)^{\otimes n_2}$$

is proportional to some element of this basis (with a proportionality constant of 0 *being allowed).*

Proof. Let S be the stabilizer group of $|\alpha_1\rangle$. There exists an isotropic subspace $\hat{S} \subset \mathbb{F}_d^{2n_1}$ and a function $f_S : \hat{S} \to \mathbb{C}$ such that

$$S = \{f_S(\hat{s})\, w(\hat{s}) \,|\, \hat{s} \in \hat{S}\}.$$

Let χ_i be the character associated with $|\alpha_i\rangle$ as in (3.71). Analogously, write the stabilizer group of $|\psi\rangle$ as

$$T = \{f_T(\hat{t})\, w(\hat{t}) \,|\, \hat{t} \in \hat{T}\}$$

for a suitable isotropic subspace $\hat{T} \subset \mathbb{F}_d^{2(n_1+n_2)}$ and phase function $f_T : \hat{T} \to \mathbb{C}$. Let

$$\hat{U} = \{\hat{t}_2 \in \mathbb{F}_d^{2n_2} \,|\, \exists \hat{s} \in \hat{S}, \hat{s} \oplus \hat{t}_2 \in \hat{T}\}.$$

The fact that $\hat{S}$ and $\hat{T}$ are isotropic implies that the same is true for $\hat{U}$.

Using Eq. (3.6), we obtain

$$\begin{aligned}
|\beta_i\rangle \langle\beta_i| &= (\langle\alpha_i| \otimes \mathrm{Id})|\psi\rangle\langle\psi|(|\alpha_i\rangle \otimes \mathrm{Id}) \\
&\propto \sum_{s\in S, t\in T} \chi_i(s) \operatorname{Tr}_1((s \otimes \mathrm{Id})t) \\
&= \sum_{\hat{s}\in\hat{S}, \hat{t}_1\oplus\hat{t}_2\in\hat{T}} \chi_i(\hat{s}) f_S(\hat{s}) f_T(\hat{t}) \operatorname{Tr}(w(\hat{s})w(\hat{t}_1))w(\hat{t}_2) \\
&\propto \sum_{\hat{t}_2\in\hat{U}} w(\hat{t}_2)\Big(\sum_{\hat{s}\in\hat{S}:\hat{s}\oplus\hat{u}\in T} \chi_i(\hat{s}) f_S(\hat{s}) f_T(\hat{s}\oplus\hat{t}_2)\Big).
\end{aligned}$$

The statement follows by invoking Lemma 3.A.1. □

Next, we show a property of stabilizer code projectors used in the proof of Lemma 3.4.8.

Lemma 3.A.3. *Suppose that P is a projector onto a stabilizer code. If P is non-diagonal, then there are $|x\rangle \neq |y\rangle$ such that $P|x\rangle = P|y\rangle \neq 0$.*

Proof. We can assume that the stabilizer group of P has $l \geq 1$ non-diagonal generators, given as $\omega^{s_j} w(z_j, x_j)$ for $s_j \in \mathbb{F}_d$, $z_j, x_j \in \mathbb{F}_d^n$, and $x_j \neq 0$. The remaining, diagonal generators are of the form $\omega^{-z\cdot b} Z(z)$ for some $b \in \mathbb{F}_d^n$ and z is an element of a suitable subspace M, such that

$$M \subset L := \{z \in \mathbb{F}_d^n \mid Z(z)w(z_j, x_j) = w(z_j, x_j)Z(z) \Leftrightarrow z \cdot x_j = 0 \quad \forall j = 1, \ldots, l\}.$$

For any $|x\rangle$, its stabilizers are $\omega^{-z\cdot x} Z(z)$ for $z \in \mathbb{F}_d^n = L \oplus L^\perp$. Under the projection P, the stabilizers with $z \in L^\perp$ are replaced by the group generated by $\omega^{s_j} w(z_j, x_j)$. For $P|x\rangle$ to be non-zero it is then necessary and sufficient that $z \cdot x = z \cdot b$ for all $z \in M$. We then have $P|x\rangle = P|y\rangle \neq 0$ if moreover $z \cdot x = z \cdot y$ for all $z \in L$. Since this enforces $\dim L = n - l$ constraints on x, there are $d^l > 1$ possible solutions. Thus, we can always find at least two distinct states $x \neq y$ such that $P|x\rangle = P|y\rangle \neq 0$, as claimed. □

3.B Properties of the counter-example Λ

In this section, we analyse the properties of the Λ-channel in the case of qubits.[2] Its action on the computational basis is given by

$$\Lambda(\rho) := \rho_{00} |+\rangle\langle+| + \sum_{x\neq 0} \rho_{xx} |x\rangle\langle x| + \frac{1}{2} \sum_{\substack{x\neq y \\ x\neq 0\neq y}} \rho_{xy} |x\rangle\langle y|, \qquad \rho_{xy} := \langle x| \rho |y\rangle. \tag{3.72}$$

From the definition, it is evident that Λ is trace-preserving. However, it is not obvious that Λ is completely stabilizer-preserving, a fact which is proven by Lem. 3.4.11. Here, we give an independent, self-contained proof for the CSP property which also sheds a bit of light on the interpretation of the channel Λ.

To this end, we claim that Λ has a Kraus decomposition given by

$$\Lambda(\rho) = H^{\otimes n} |0\rangle\langle 0| \rho |0\rangle\langle 0| H^{\otimes n} + \frac{1}{2^{n-1}} \sum_{z\in\mathbb{F}_2^n\setminus 0} P_z \rho P_z, \tag{3.73}$$

where $P_z = (\mathbb{1} - Z(z))/2$ projects onto the stabilizer code given by the span of computational basis states $|x\rangle$ with $x \cdot z \neq 0$. Then,

[2] A similar analysis can also be done for qudits which is however more evolved.

by the polar decomposition, Eq. (3.69), of CSP channels discussed in App. 3.5, the Kraus decomposition (3.73) defines a CSP channel since

$$|0\rangle\langle 0| + \frac{1}{2^n} \sum_{z \in \mathbb{F}_2^n \setminus 0} (\mathbb{1} - Z(z)) = \mathbb{1} + |0\rangle\langle 0| - \frac{1}{2^n} \sum_{z \in \mathbb{F}_2^n} Z(z) = \mathbb{1}. \quad (3.74)$$

Alternatively, it is also straightforward to compute the Choi state from Eq. (3.73). Let us define for any $z \in \mathbb{F}_2^n \setminus 0$ the affine subspace $K_z := \{x \in \mathbb{F}_2^n : z \cdot x = 1\}$ and the $2n$-qubit stabilizer state

$$|\psi_z\rangle := 2^{-\frac{n-1}{2}} \sum_{x \in K_z} |xx\rangle . \quad (3.75)$$

Then, the Choi state is

$$\mathcal{J}(\Lambda) = \frac{1}{2^n} \left(|+\rangle\langle +| \otimes |0\rangle\langle 0| + \sum_{z \neq 0} |\psi_z\rangle\langle \psi_z| \right), \quad (3.76)$$

which lies in the stabilizer polytope SP_{2n}.

Finally, to prove the Kraus decomposition (3.73), we check that it agrees with Eq. (3.72) on the computational basis. To this end, let us denote the channel Eq. (3.73) as $\tilde{\Lambda}$. Note that $P_z |x\rangle\langle y| P_z$ is zero if and only if x or y is orthogonal to z and $|x\rangle\langle y|$ otherwise. Thus, $\tilde{\Lambda}(|0\rangle\langle 0|) = |+\rangle\langle +|$. For any $x \neq 0$, the linear equation $x \cdot z = 1$ has exactly 2^{n-1} solutions $z \in \mathbb{F}_2^n$. Since the first term in Eq. (3.73) yields 0, we get $\tilde{\Lambda}(|x\rangle\langle x|) = |x\rangle\langle x|$ for any $x \neq 0$. Furthermore, adding the condition $y \cdot z = 1$ for any $y \notin \{0, x\}$ will further half the solution space, yielding 2^{n-2} vectors which are not orthogonal to both x and y. Thus, given two non-zero vectors $x \neq y$, we get $\tilde{\Lambda}(|x\rangle\langle y|) = \frac{1}{2} |x\rangle\langle y|$ which then shows that $\tilde{\Lambda} = \Lambda$.

A natural question to ask is whether Λ can be expressed in terms of more elementary quantum channels. We can write the channel as a composition of the following three operations:

1. Perform a projective measurement with projectors $\{|0^n\rangle\langle 0^n|, \mathbb{1} - |0^n\rangle\langle 0^n|\}$. This channel sets all off-diagonal terms in the first row and column of ρ to zero, i.e. it block-diagonalises ρ with respect to the entry at position $(0, 0)$.

2. Partial dephasing in the computational basis with probability 1/2. This channel reduces the amplitude of the off-diagonal terms by 1/2.

3. Apply a global Hadamard gate on all qubits conditioned on the "0" outcome of the measurement.

Interestingly, all three components are necessary for Λ to have the desired properties. If we leave out the second channel, it is possible to show that the composition of 1 and 3 is not stabilizer-preserving for $n \geq 2$[3], while for $n = 1$ it is simply a stabilizer operation. Moreover, if we leave out channel 2 and 3, then we can rewrite the block-diagonalisation as a uniform convex combination of the identity and the diagonal n-qubit gate $V_n := \mathrm{diag}(-1, 1, \ldots, 1)$. Note that $V_n = X^{\otimes n}(C^{n-1}Z)X^{\otimes n}$, thus it is in the n-th level of the Clifford hierarchy. Hence, for $n \leq 2$, this is a mixed Clifford channel and in particular a stabilizer operation. For $n > 2$, the same technique as before can be used to show that this channel is not CSP. The effect of the dephasing channel is to sufficiently reduce the "magic" of the overall channel. With increasing dephasing strength, it approaches the CSP polytope from the outside and eventually becomes CSP. Figuratively speaking, the Hadamard gate in the last step fine-tunes the direction from which the CSP polytope is being approached, resulting in a channel which is a vertex.

3.C Measurements that are not followed by adaptive operations are never extremal

Lemma 3.C.1. *Suppose $\mathcal{E} = \sum_i K_i \cdot K_i \in \mathrm{CSP}_n$ is an extremal CSP map with Kraus operators K_i. Then $\mathcal{E}$ does not contain a set of d Kraus operators that are Pauli measurements of the form $PP_0, ..., PP_{d-1}$ for some fixed Pauli projector P and where $P_0, ..., P_{d-1}$ are projectors onto the d eigenspaces of some Pauli operator $w(z, x)$.*

[3]This can in principle be done by computing the Choi states of the corresponding channels and then finding a hyperplane that separates them from the stabilizer polytope SP_{2n}

Proof. Let $\mathcal{O}$ be the map that is composed of the d Kraus operators $PP_0, ..., PP_{d-1}$, i.e.

$$\mathcal{O}(\rho) = \sum_{x=0}^{d-1} PP_x \rho P_x P.$$

There is a Clifford C such that $PP_x = C\tilde{P} \otimes |x\rangle\langle x| C^\dagger$ for a projector $\tilde{P}$ acting on the first $n-1$ qudits. Define the operation

$$\mathcal{M}(\rho) = \sum_{x=0}^{d-1} \tilde{P} \otimes |x\rangle\langle x| \rho \tilde{P} \otimes |x\rangle\langle x|, \tag{3.77}$$

so $\mathcal{O}(\rho) = C\,\mathcal{M}(C^\dagger \rho C)\, C^\dagger$.

Let $s_0, ..., s_{d-1}$ be the eigenstates of some Pauli operator $w(z,x)$ for $z, x \in \mathbb{F}_1$ and let $C_i = \mathrm{diag}(ds_i) \in \mathrm{Cl}_1$ be the diagonal Clifford unitary with diagonal proportional to s_i. We claim that

$$\mathcal{M}(\rho) = \frac{1}{d} \sum_{i=0}^{d-1} (\tilde{P} \otimes C_i) \rho (\tilde{P} \otimes C_i^\dagger). \tag{3.78}$$

It suffices to check the equation for inputs of the form $A \otimes |x\rangle\langle y|$ for $x, y \in \mathbb{F}_d$ and a Hermitian matrix A acting on $n-1$ qubits. We have $\mathcal{M}(A \otimes |x\rangle\langle y|) = \langle x|y\rangle\, \tilde{P}A\tilde{P} \otimes |x\rangle\langle y|$ and for the RHS of (3.78)

$$\frac{1}{d} \sum_{i=0}^{d-1} \tilde{P}A\tilde{P} \otimes C_i |x\rangle\langle y| C_i^\dagger = \frac{1}{d} \tilde{P}A\tilde{P} \otimes \sum_{i=0}^{d-1} C_i |x\rangle\langle y| C_i^\dagger = \tilde{P}A\tilde{P} \otimes \sum_{i=0}^{d-1} s_i(x)\overline{s_i(y)}\, |x\rangle\langle y| \tag{3.79}$$

$$= \tilde{P}A\tilde{P} \otimes |x\rangle\langle y| \sum_{i=0}^{d-1} s_i(x)\overline{s_i(y)} = \langle x|y\rangle\, \tilde{P}A\tilde{P} \otimes |x\rangle\langle y|, \tag{3.80}$$

where the last equality stems from the fact that

$$\sum_{i=0}^{d-1} s_i(x)\overline{s_i(y)} = \Big(\sum_{i=0}^{d-1} |s_i\rangle\langle s_i| \Big)(x,y) = \mathbb{1}_d(x,y) = \langle x|y\rangle\, |x\rangle\langle y|. \tag{3.81}$$

Hence,

$$\mathcal{O}(\rho) = C\,\mathcal{M}(C\rho C^\dagger)\, C^\dagger = CM(C^\dagger \rho C)\, C^\dagger \tag{3.82}$$

$$= \frac{1}{d} \sum_{i=0}^{d-1} C(\tilde{P} \otimes C_i) C^\dagger \rho\, C(\tilde{P} \otimes C_i) C^\dagger. \tag{3.83}$$

If the original channel $\mathcal{E}$ decomposes as $\mathcal{E} = \mathcal{O} + \mathcal{O}^c$, then we can write it now as a convex combination of distinct operations

$$\mathcal{E} = \frac{1}{d}(\mathcal{E}_1 + \cdots + \mathcal{E}_{d-1}) \quad \text{with} \quad \mathcal{E}_i(\rho) = C(\tilde{P} \otimes C_i)C^\dagger \, \rho \, C(\tilde{P} \otimes C_i)C^\dagger + \mathcal{O}^c(\rho).$$

The maps $\mathcal{E}_i$ are completely positive and trace-preserving because

$$\begin{aligned}
\mathcal{E}_i^\dagger(\mathbb{1}_n) &= C(\tilde{P} \otimes C_i^\dagger)C^\dagger \mathbb{1}_n C(\tilde{P} \otimes C_i)C^\dagger + (\mathcal{O}^c)^\dagger(\mathbb{1}_n) \\
&= C(\tilde{P} \otimes \mathbb{1}_1)C^\dagger + (\mathcal{O}^c)^\dagger(\mathbb{1}_n) \\
&= \mathcal{O}^\dagger(\mathbb{1}_n) + (\mathcal{O}^c)^\dagger(\mathbb{1}_n) \\
&= \mathcal{E}^\dagger(\mathbb{1}_n) \\
&= \mathbb{1}_n,
\end{aligned}$$

where the third equation follows from

$$\begin{aligned}
\mathcal{O}^\dagger(\mathbb{1}_n) &= \sum_{x=0}^{d-1} C(\tilde{P} \otimes |x\rangle\langle x|)C^\dagger \mathbb{1}_n C(\tilde{P} \otimes |x\rangle\langle x|)C^\dagger \\
&= C\Big(\sum_{x=0}^{d-1} \tilde{P} \otimes |x\rangle\langle x| \Big)C^\dagger \\
&= C(\tilde{P} \otimes \mathbb{1}_1)C^\dagger.
\end{aligned}$$

This proves that $\mathcal{E}$ cannot be extremal. □

Chapter 4

About the Λ-polytope

About this section

This section has not been published. The author of this thesis is the only contributor to the results presented in this section. Section 4.4 arose from discussions with Michael Zurel, Cihan Okay and Robert Raussendorf.

4.1 Introduction and Summary of results

The Λ-polytope has gained increasing attention in recent years [RBVT$^+$20, Zur20, ZOR20, ZORH21, Hei19, OZR21, OHG]. Using the concept of polar duality from polyhedral geometry, the Λ-polytope is precisely the polar dual stabilizer polytope and its vertices define the facets of the stabilizer polytope. Hence, studying Λ might help to gain new insights about geometric properties of stabilizer states.

On the other hand, again by polar duality, the polytope Λ contains the set of density matrices, i.e. all quantum states. Moreover, Λ is also deeply connected to the stabilizer formalism, which is considered as a classical subtheory of quantum computation. Thus, one might ask:

What should be considered classical within Λ and what quantum?

This question is also strongly motivated by recent work of Zurel, Raussendorf and Okay [ZOR20]. They developed a classical simulation algorithm for quantum computation with magic states (QCM), which is based on the Λ-polytope in the following sense: The idea

of the algorithm is to express a quantum state as a convex combination of vertices in Λ, then to sample a vertex from the distribution induced by the expansion coefficients of the convex combination, and then to update the sampled vertex under Pauli measurements and Clifford unitaries. This algorithm produces the same outcomes as the corresponding quantum algorithm [1]. The crucial observation is that the polytope Λ is stable under Clifford unitaries and Pauli measurements [ZOR20]. This means that acting on an element in Λ by a Clifford unitary or Pauli measurements returns another element in Λ.

Unfortunately, obtaining all vertices of the Λ-polytope for an arbitrary number of qubits or qudits seems to be an extremely hard task. Complete descriptions of Λ via its vertices are only known for low dimensions (one or two qubits, one qudit, when d is an odd prime) [Hei19, CGG$^+$06, Rei05, OZR21].

In this chapter, we will deal with this question for the case of odd prime dimensional systems. We will classify a particular class of operators that live on the boundary of Λ. Their common feature is that, when expanded in the generalized Pauli basis, their coefficients are either zero or complex roots of unity. This family of operators encompasses stabilizer states and Wigner operators, where the latter have been shown to be vertices of Λ [VFGE12, ZORH21]. Furthermore, the qubit analogue of this family (here the expansion coefficients in the Pauli basis are $0, \pm 1$) also contains vertices of the qubit Λ-polytope [Hei19, OZR21].

Our findings will show that for odd prime dimensional systems, this family of operators splits up into two parts:
Either they are of Wigner type (the underlying support in the generalized Pauli basis is a subspace) or they are cnc (as defined in [RBVT$^+$20], cnc refers to "closed under inference and non-contextual"). Conjecturally, there is a subfamily of qudit cnc-type operators that are vertices of Λ, precisely those with inclusion maximal support in the generalized Pauli basis. This conjecture is supported by numerical experiments for two qudits (where the underlying dimension is three) [Zur21].

Furthermore, we will argue why these operators should be con-

[1] Observe that we do not make any statements about efficiency here.

sidered as classical in QCM: We will briefly sketch why they can be efficiently classically updated under Clifford unitaries and Pauli measurements.

The remaining part of this chapter is organized as follows: In Section 4.2, we give an introduction to the relevant concepts used throughout the main part of this chapter. We state our main result prove it in Section 4.3. In Section 4.4 we sketch how to classically compute updates of this new class of operators under Clifford unitaries and Pauli measurements.

4.2 Preliminaries

We use the notation for the stabilizer formalism, as introduced in Section 3.2.1 and only consider the case where the dimension d is an odd prime. Let STAB_n be the set of n-qudit stabilizer states, viewed as rank-1 density matrices. The Λ-polytope, is given by

$$\Lambda_n := \{X \in \mathrm{Herm}_1(d^n) \mid \mathrm{Tr}(SX) \geq 0 \text{ for all } S \in \mathrm{STAB}_n\}, \quad (4.1)$$

where $\mathrm{Herm}_1(d^n)$ is the set of $d^n \times d^n$ Hermitian matrices of trace one. The set Λ_n is a polytope for any number of qudits n and any dimension d; even in the case where d is not prime, see [ZORH21, Lemma 1]. Using the concept of polarity from polyhedral geometry [Zie95, Section 2.3], it corresponds to the polar dual stabilizer polytope (see Appendix C of [ZORH21] for further details). Polar duality gives a one-to-one correspondence between vertices of Λ_n and facets of the stabilizer polytope. Its qubit version was studied in [Hei19, ZOR20, OZR21].

We want to characterize Hermitian operators in Λ_n whose expansion coefficients in the generalized Pauli basis are roots of unity. Recall that the generalized Pauli basis is given by

$$w(u) := \omega^{(u_Z^\mathsf{T} u_X)/2} Z(u_z) X(u_x), \quad \omega = e^{2\pi i/d}, \quad (4.2)$$

where $u = (u_Z, u_X) \in \mathbb{F}_d^n \times \mathbb{F}_d^n \cong \mathbb{F}_d^{2n}$. As usual, we will consider $\mathbb{F}_d^{2n}$ as a symplectic vector space with symplectic inner product

$$[u, v] := u_Z^\mathsf{T} v_X - u_X^\mathsf{T} v_Z.$$

Our goal is to study Hermitian operators of the form

$$A_\Omega^\eta = \frac{1}{d^n} \sum_{u \in \Omega} \omega^{\eta(u)} w(u), \qquad \Omega \subset \mathbb{F}_d^{2n},\ \eta : \Omega \to \mathbb{R} \tag{4.3}$$

that are contained in Λ_n. Here, we allow η take values in $\mathbb{R}$. In the sequel we will often have the situation that $\eta : \Omega \to \mathbb{F}_d$. When computing $\omega^{\eta(u)}$, we interpret $\eta(u)$ as the corresponding integer in $\mathbb{Z}$. Note that this class of operators encompasses stabilizer states (Ω is a Lagrangian subspace, i.e. an isotropic subspace of dimension n and $\eta : \Omega \to \mathbb{F}_d$ linear) and Wigner operators ($\Omega = \mathbb{F}_d^n$ and $\eta : \Omega \to \mathbb{F}_d$ is linear). The latter have been shown to be vertices of Λ_n for any odd dimension [VFGE12, ZORH21]. In Corollary 4.3.5, we will show that if $\Omega \neq 0$ and $A_\Omega^\eta \in \Lambda_n$, then A_Ω^η will be at least contained in a non-trivial face of Λ_n.

4.3 Main result

We will show that containment of A_Ω^η in Λ_n for A_Ω^η as in (4.3) imposes restrictions on the set Ω and the function η. For a linear subspace $L \subset \mathbb{F}_d^{2n}$ let

$$L^* := \{\gamma : L \to \mathbb{F}_d,\ \gamma(a+b) = \gamma(a) + \gamma(b)\}$$

be its dual space, i.e. the space of linear functions on L. Further, let

$$L^\perp = \{a \in \mathbb{F}_d^{2n} \,:\, [a,b] = 0 \text{ for all } b \in L\}$$

be the orthogonal complement of L. If $L \subset L^\perp$, then L is called isotropic and if $L = L^\perp$, then L is Lagrangian. For $\eta : \Omega \to \mathbb{R}$, we denote the restriction of η to some subset $K \subset \Omega$ by $\eta_{|K}$ and by $\langle K \rangle$ we denote the group generated by the elements of K. The restrictions on Ω and η for $A_\Omega^\eta \in \Lambda_n$ are summarized in the following theorem:

Theorem 4.3.1. *If $A_\Omega^\eta \in \Lambda_n$, then*

(i) if $I \subset \Omega$ is an isotropic subspace, then $\eta_{|I} \equiv \gamma \pmod d$ for some $\gamma \in I^*$,

(ii) the set Ω is closed under addition of orthogonal elements, i.e., if $a, b \in \Omega$ and $[a, b] = 0$, then $a + b \in \Omega$.

Operators with η as in (i) and Ω is as in (ii) are usually referred to as *cnc-operators* [RBVT$^+$20, KL19], meaning *closed under inference and non-contextual*. Closed under inference expresses that Ω satisfies property (ii) of the theorem and non-contextual means that η defines a non-contextual value assignment on Ω; see [DOBV$^+$17, Definition 1].

As a consequence, if $A^\eta_\Omega \in \Lambda_n$, then we can view η as a function $\eta : \Omega \to \mathbb{F}_d$. Classifying all operators $A^\eta_\Omega \in \Lambda_n$ requires a characterization of all sets $\Omega \subset \mathbb{F}_d^{2n}$ that are closed under inference. Together with Theorem 4.3.1, we obtain:

Theorem 4.3.2. *If A^η_Ω is of the form* (4.3)*, then $A^\eta_\Omega \in \Lambda_n$ if and only if*

(i) Ω is a subspace of $\mathbb{F}_d^{2n}$ and $\eta \in \Omega^*$ or

(ii) Ω is of the form

$$\Omega = \langle I, h_1\rangle \cup \langle I, h_2\rangle \cup \cdots \cup \langle I, h_\ell\rangle, \tag{4.4}$$

where $[h_i, h_j] \neq 0$ for all $i \neq j$ and I is an isotropic subspace with $h_i \in I^\perp$ for $i = 1, \ldots, \ell$. In addition, $\eta_{|\langle h_i, I\rangle} \in \langle h_i, I\rangle^*$. The function η is uniquely determined by $\eta_{|I} \in I^*$ and $\eta(h_1), \ldots, \eta(h_\ell) \in \mathbb{F}_d$.

4.3.1 Comparison to qubits

An analogous classification has been done for qubits, i.e. $d = 2$. In this case, the operators of interest are given by

$$A^\gamma_\Omega = \frac{1}{2^n} \sum_{u \in \Omega} (-1)^{\gamma(u)} w(u), \quad \Omega \subset \mathbb{F}_2^{2n}, \gamma : \Omega \to \mathbb{F}_2,$$

where $w(u)$ are the typical Pauli matrices. Let $\beta : \mathbb{F}_2^{2n} \times \mathbb{F}_2^{2n} \to \mathbb{F}_2$ be such that the composition law of the (qubit) Pauli matrices is given by $(-1)^{\beta(a,b)} w(a)w(b) = w(a + b)$. For qubits, we have the following classification:

Theorem 4.3.3 ([Hei19, OZR21])**.** *An Hermitian operator A_Ω^γ is contained in Λ_n if and only if Ω is of the form*

$$\Omega = \langle I, h_1\rangle \cup \langle I, h_2\rangle \cup \cdots \cup \langle I, h_\ell\rangle, \tag{4.5}$$

where $[h_i, h_j] \neq 0$ for all $i \neq j$ and I is an isotropic subspace with $h_i \in I^\perp$ for $i = 1, \ldots, \ell$ and for any $a, b \in \Omega$ with $[a, b] = 0$ we have

$$\gamma(a) + \gamma(b) + \beta(a, b) = \gamma(a + b). \tag{4.6}$$

Furthermore, if Ω is inclusion maximal, i.e. there is no Ω' of the form (4.5) that strictly contains Ω, then A_Ω^γ is a vertex of Λ_n.

Compared to the case of odd-prime dimensional systems we have two important differences. First, on every isotropic subspace $\langle h_i, I\rangle \subset \Omega$ the function $\gamma_{|\langle h_i, I\rangle}$ is only linear up the shift induced by β. Second, if Ω is a non-isotropic subspace, one can construct a Mermin-square within in Ω, which does not allow a value assignment that satisfies (4.6) [Hei19, RBVT$^+$20]. This implies that $A_\Omega^\gamma \notin \Lambda_n$ for all $\gamma : \Omega \to \mathbb{F}_2$ whenever Ω contains a non-isotropic subspace.

4.3.2 Proof of Theorem 4.3.1

We will proceed with the proof of Theorem 4.3.1. For $M \subset \mathbb{F}_d^{2n}$ let $\mathrm{pr}_M : \mathrm{Herm}(d^n) \to \mathrm{span}\{w(b) : b \in M\}$ be the projection that acts via

$$\sum_{b \in \mathbb{F}_d^{2n}} c_b w(b) \overset{\mathrm{pr}_M}{\longmapsto} \sum_{b \in M} c_b w(b).$$

To prove Theorem 4.3.1, we will use the fact that if $X \in \Lambda_n$, then $\mathrm{pr}_M(X) \in \mathrm{pr}_M(\Lambda_n)$. If $I \subset \mathbb{F}_d^{2n}$ is an isotropic subspace and $\gamma \in I^*$ we will fix the notation

$$\Pi_I^\gamma = A_I^\gamma = \frac{1}{d^n} \sum_{u \in I} \omega^{\gamma(u)} w(u). \tag{4.7}$$

Note that Π_I^γ is proportional to the projector onto an $[n, n - \dim(I)]$ stabilizer code [NC11, Proposition 10.5]. Lemma 9 in [ZORH21] allows

us to add redundant inequalities to the description of Λ_n, so that we can write

$$\Lambda_n = \{X \in \mathrm{Herm}_1(d^n) \mid \mathrm{Tr}(\Pi_I^\gamma X) \geq 0 \text{ for all isotropic subspaces } I,\ \gamma \in I^*\}. \tag{4.8}$$

We will be mainly interested in $\mathrm{pr}_I(\Lambda_n)$ for I being an isotropic subspace.

Lemma 4.3.4. *If I is an isotropic subspace, then $\mathrm{pr}_I(\Lambda_n)$ is a self dual simplex with vertices $\Pi_I^\gamma, \gamma \in I^*$. Hence,*

$$\mathrm{pr}_I(\Lambda_n) = \mathrm{conv}\{\Pi_I^\gamma \,|\, \gamma \in I^*\}. \tag{4.9}$$

Proof. Consider the affine subspace $U_I = \{\sum_{b\in I} c_b w(b) \,:\, c_b \in \mathbb{C}\} \cap \mathrm{Herm}_1(d^n)$. Due to $\mathrm{pr}_I(\Pi_I^\gamma) = \Pi_I^\gamma$ and the description of Λ_n in (4.8), it follows that

$$\mathrm{pr}_I(\Lambda_n) \subseteq \{X \in U_I \mid \mathrm{Tr}(X\Pi_I^\gamma) \geq 0 \text{ for all } \gamma \in I^*\}. \tag{4.10}$$

On the other hand, character orthogonality gives

$$\mathrm{Tr}(\mathrm{pr}_I(\Pi_I^\gamma)\mathrm{pr}_I(\Pi_I^{\tilde\gamma})) = \mathrm{Tr}(\Pi_I^\gamma \Pi_I^{\tilde\gamma}) = \delta_{\gamma=\tilde\gamma}\frac{|I|}{d^n} \tag{4.11}$$

for all $\gamma, \tilde\gamma \in I^*$, so the operators $\Pi_I^\gamma, \gamma \in I^*$ are $|I|$ affinely independent points in U_I and $\mathrm{conv}\{\Pi_I^\gamma \,|\, \gamma \in I^*\}$ is a simplex with $|I|$ facets. Again, by (4.11), the facet normals are given by Π_I^γ, implying that this simplex is self dual. In summary,

$$\mathrm{pr}_I(\Lambda_n) \subseteq \{X \in U_I \mid \mathrm{Tr}(X\Pi_I^\gamma) \geq 0 \text{ for all } \gamma \in I^*\} = \mathrm{conv}\{\Pi_I^\gamma \,|\, \gamma \in I^*\} \subseteq \mathrm{pr}_I(\Lambda_n),$$

which gives the desired result. □

We will use the characterization of $\mathrm{pr}_I(\Lambda_n)$ for I being an isotropic subspace to prove Theorem 4.3.1. The overall strategy will be to argue that if Ω and η violate one of the conditions of the theorem, then there is an isotropic subspace $I \subset \mathbb{F}_d^{2n}$ such that $\mathrm{pr}_I(A_\Omega^\gamma) \notin \mathrm{pr}_I(\Lambda_n)$, implying $A_\Omega^\gamma \notin \Lambda_n$. To show $\mathrm{pr}_I(A_\Omega^\gamma) \notin \mathrm{pr}_I(\Lambda_n)$, we will construct a hyperplane that separates the point $\mathrm{pr}_I(A_\Omega^\gamma)$ from the projected polytope $\mathrm{pr}_I(\Lambda_n)$. That is, we construct some explicit Hermitian operator

$Y \in \operatorname{span}\{w(b) : b \in I\}$ such that

$$\operatorname{Tr}(\operatorname{pr}_I(A_\Omega^\gamma)Y) > \max_{X \in \operatorname{pr}_I(\Lambda_n)} \operatorname{Tr}(XY) = \max_{\gamma \in I^*} \operatorname{Tr}(\Pi_I^\gamma Y), \tag{4.12}$$

where the last equation is a consequence of Lemma 4.3.4.

Proof of Theorem 4.3.1. Let $A_\Omega^\eta \in \operatorname{Herm}_1(d^n)$ as in (4.3) with $\eta : \Omega \to \mathbb{R}$. As A_Ω^η is Hermitian, we have

$$\frac{1}{d^n} \sum_{u \in \Omega} \omega^{\eta(u)} w(u) = \frac{1}{d^n} A_\Omega^\eta = (A_\Omega^\eta)^\dagger = \sum_{u \in \Omega} \omega^{-\eta(u)} T_{-u},$$

implying $-\eta(u) = \eta(-u)$ for all $u \in \Omega$.

(i) Let I be an isotropic subspace contained in Ω. We will show that if $A_\Omega^\eta \in \Lambda_n$, then $\eta_{|I} \equiv \gamma$ for some linear function $\gamma \in I^*$. Therefore, assume that $\eta_{|I} \not\equiv \gamma$ for all $\gamma \in I^*$. In the sense of Equation (4.12), we can separate $\operatorname{pr}_I(\Lambda)$ and $\operatorname{pr}_I(A_\Omega^\eta)$ by the hyperplane with normal vector $A_I^{\eta_{|I}} \in \operatorname{span}\{w(b) : b \in I\}$:

$$\operatorname{Tr}(A_I^{\eta_{|I}} \operatorname{pr}_I(A_\Omega^\eta)) = \operatorname{Tr}(A_I^{\eta_{|I}} A_I^{\eta_{|I}}) = \frac{1}{d^{2n}} \sum_{u \in I} \omega^{\eta(u)-\eta(u)} \operatorname{Tr}(w(u)w(-u)) = \frac{|I|}{d^n} > \operatorname{Tr}(A_I^{\eta_{|I}} \Pi$$

where the strict inequality holds for all $\gamma : I \to \mathbb{F}_d$ with $-\gamma(u) = \gamma(-u)$ and $\gamma \neq \eta_{|I}$, so in particular for all $\gamma \in I^*$.

(ii) Now suppose that $\eta_{|I} \in I^*$ for any isotropic subspace I contained in Ω. For (ii) assume that there are $a, b \in \Omega$ with $[a, b] = 0$ such that $a+b \notin \Omega$. We may assume that $b \neq -a$ since $0 \in \Omega$, due to $\operatorname{Tr}(A_\Omega^\eta) = 1$.

Set $I = \langle a, b \rangle$, so consequently $I \not\subseteq \Omega$. Define the Hermitian matrix $Y = \mathcal{A}_\mathcal{M}^\nu$ with

$$\begin{aligned} \mathcal{M} &= \{\pm a, \pm b, \pm(a+b)\}, \\ \nu : \mathcal{M} \to \mathbb{R}, \quad \nu(a) &= \eta(a),\ \nu(b) = \eta(b),\ \nu(a+b) = \eta(a) + \eta(b) + d/2, \\ \nu(-x) &= -\nu(x) \text{ for all } x \in \mathcal{M}. \end{aligned}$$

Then

$$\operatorname{Tr}(A_\Omega^\gamma Y) = \frac{|\mathcal{M} \cap \Omega|}{d^n} = \frac{4}{d^n}$$

and for all $\gamma \in I^*$

$$\begin{aligned} d^n \operatorname{Tr}(\Pi_I^\gamma Y) &= \left(\omega^{\eta(a)-\gamma(a)} + \omega^{-\eta(a)+\gamma(a)}\right) + \left(\omega^{\eta(b)-\gamma(b)} + \omega^{-\eta(b)+\gamma(b)}\right) \\ &\quad + \left(\omega^{\eta(a)+\eta(b)-\gamma(a)-\gamma(b)+d/2} + \omega^{-(\eta(a)+\eta(b)-\gamma(a)-\gamma(b)+d/2)}\right) \\ &= 2\cos\left(\frac{2\pi(\eta(a)-\gamma(a))}{d}\right) + 2\cos\left(\frac{2\pi(\eta(b)-\gamma(b))}{d}\right) \\ &\quad + 2\cos\left(\frac{2\pi\big(\eta(a)-\gamma(a)+\eta(b)-\gamma(b)+d/2\big)}{d}\right) \\ &= 2\left(\cos(x)+\cos(y)+\cos(x+y+\pi)\right) \\ &= 2\left(\cos(x)+\cos(y)-\cos(x+y)\right), \end{aligned}$$

where

$$x = \frac{2\pi(\eta(a)-\gamma(a))}{d} \quad \text{and} \quad y = \frac{2\pi(\eta(b)-\gamma(b))}{d}$$

and where we used the linearity of γ in the first equality. Further, observe that

$$\cos(a)+\cos(b)-\cos(a+b) \leq 3/2 \quad \text{for all } a, b \in \mathbb{R}.$$

For example, this can be seen by computing the local maxima of the given function. As the function is periodic, the local maxima give upper bounds for the values that the function can attain. Consequently,

$$\operatorname{Tr}(\Pi_I^\gamma Y) \leq \frac{3}{d^n} < \frac{4}{d^n} = \operatorname{Tr}(A_\Omega^\gamma Y),$$

for all $\gamma \in I^*$, implying $\operatorname{pr}_I(A_\Omega^\gamma) \notin \operatorname{pr}_I(\Lambda_n)$ and $A_\Omega^\gamma \notin \Lambda_n$. □

Additionally, Theorem 4.3.1 has the following consequence:

Corollary 4.3.5. *Assume that $\Omega \neq 0$ and $A_\Omega^\eta \in \Lambda_n$. Then $A_\Omega^\eta \in \Lambda_n$ lies on the boundary of Λ_n.*

Proof. Since $\Omega \neq 0$, Theorem 4.3.1 implies that there is $a \in \Omega$ such that the 1-dimensional isotropic subspace $I = \langle a \rangle$ is contained in Ω and $\eta_{|I} \in I^*$. Now, using Lemma 4.3.4, we obtain

$$\begin{aligned} \max_{X \in \Lambda_n} \operatorname{Tr}(X\Pi_I^{\eta_{|I}}) = \max_{X \in \operatorname{pr}_I(\Lambda_n)} \operatorname{Tr}(X\Pi_I^{\eta_{|I}})) &= \max_{\gamma \in I^*} \operatorname{Tr}(\Pi_I^\gamma \Pi_I^{\eta_{|I}}) \\ &= \operatorname{Tr}((\Pi_I^{\eta_{|I}})^2) = \operatorname{Tr}(A_\Omega^\eta \Pi_I^\eta). \end{aligned}$$

Hence, A_Ω^η lies on the hyperplane

$$\{X \in \mathrm{Herm}_1(d^n) \,:\, \mathrm{Tr}(X\Pi_I^{\eta_{|I}}) = \mathrm{Tr}((\Pi_I^{\eta_{|I}})^2)\},$$

which is a supporting hyperplane of Λ_n. □

4.3.3 Classifying all sets that are closed under addition of orthogonal elements

As shown in Theorem 4.3.1, a necessary condition for $A_\Omega^\eta \in \Lambda_n$ is that Ω is closed under addition of orthogonal elements. In this section, we will characterize all sets with this property.

Proposition 4.3.6. *Any set $\Omega \subseteq \mathbb{F}_d^{2n}$ that is closed under addition of orthogonal elements is a subspace or it is of the form*

$$\Omega = \langle I, h_1\rangle \cup \langle I, h_2\rangle \cup \cdots \cup \langle I, h_\ell\rangle, \tag{4.13}$$

where $[h_i, h_j] \neq 0$ for all $i \neq j$ and I is an isotropic subspace with $h_i \in I^\perp$ for $i = 1, \ldots, \ell$.

One can easily verify that sets of the form (4.13) are indeed closed under inference, see [RBVT+20, Lemma 3] for a proof[2].

To prove the proposition, we define some necessary concepts. For an arbitrary set $\Omega \subset \mathbb{F}_d^{2n}$ let $O(\Omega) \subseteq \mathbb{F}_d^{2n}$ be the *orthogonal closure* of Ω, i.e. the smallest set which contains Ω and for all $u, v \in O(\Omega)$ with $[u, v] = 0$ we have $u + v \in O(\Omega)$. For $v \in \mathbb{F}_d^{2n}$ we define its orthogonal complement by $v^\perp = \{u \in \mathbb{F}_d^{2n} \,:\, [u, v] = 0\}$, which is a $(2n-1)$-dimensional subspace. Furthermore, for $M \subset \mathbb{F}_d^{2n}$ set $M^\times = M \setminus \{0\}$.

If $\Omega \subset \mathbb{F}_d^{2n}$, then its undirected *orthogonality graph* is the graph $G(\Omega) = (\Omega, E)$ with vertex set Ω and edge set

$$E = \{\{a, b\} \in \Omega \times \Omega \,:\, a \neq b,\, [a, b] = 0\}.$$

The overall proof strategy is induction: Start with one element $u \in \mathbb{F}_d^{2n}$. Obviously $O(\{u\}) = \langle u\rangle$, which is a subspace and simultaneously of the form (4.13). In the induction step, suppose that a set $\Omega \subset \mathbb{F}_d^{2n}$ is a subspace or of the form (4.13). Then we show that for all $v \in \mathbb{F}_d^{2n}$ the orthogonal closure $O(\Omega \cup \{v\})$ is again a subspace or of the form (4.13). This will be shown in Lemma 4.3.7, respectively Lemma 4.3.8.

[2]The proof can be straightforwardly adapted from the case $d = 2$ to d being an odd prime.

Lemma 4.3.7. *Let $\Omega \subset \mathbb{F}_d^{2n}$ be a subspace and $v \in \mathbb{F}_d^{2n}$. Then $O(\Omega \cup \{v\})$ is of the form (4.13) if and only if $\Omega \cap v^\perp$ is isotropic. Otherwise $O(\Omega \cup \{v\})$ is the subspace $\langle v, \Omega \rangle$.*

Lemma 4.3.8. *Let $\Omega \subset \mathbb{F}_d^{2n}$ be of the form (4.13) and $v \in \mathbb{F}_d^{2n}$. Then $O(\Omega \cup \{v\})$ is of the form (4.13) or a subspace.*

To prove the lemmata, we will require some auxiliary statements. One crucial observation is the following:
Suppose you are given $\Omega = \{a, b, c, d\} \in \mathbb{F}_d^{2n}$ such that the orthogonality graph $G(\Omega)$ is a 4-cycle and $\dim(\langle a, b, c, d \rangle) = 4$. If the underlying field is $\mathbb{F}_2$, then the orthogonal closure $O(\Omega)$ is the Mermin square, together with 0. In contrast, if the underlying field is $\mathbb{F}_d$ and d is an odd prime, then $O(\Omega)$ is the 4-dimensional subspace spanned by a, b, c, d; see Lemma 4.3.10.

Lemma 4.3.9. *Assume that $\Omega = \{a, b, c, d\}$ such that $\dim(\langle a, b \rangle) = \dim(\langle c, d \rangle) = 2$ and $G(\Omega)$ is given by Figure 4.1. Then $O(\Omega)$ is of the form (4.13) or it contains a subset M with $|M| = 4$ such that $G(M)$ is a 4-cycle.*

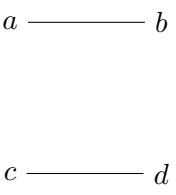


Figure 4.1: Initial orthogonality graph of Lemma 4.3.9.

Proof. Since $\dim(\langle c, d \rangle) = 2$, there is $x \in \langle c, d \rangle^\times \cap a^\perp \subset O(\Omega)$. Now we distinguish two cases:

(1) $[b, x] = 0$:

We are in the setting of Figure 4.2 (left). Then there is $y \in \langle a, x \rangle^\times \cap d^\perp \subset O(\Omega)$ and for $I = \langle x, y \rangle$ we have $b, d \in I^\perp$ (see Figure 4.2, right), Now the closure of Ω is given by

$$O(\Omega) = \langle I, b \rangle \cup \langle I, d \rangle.$$

(2) $[b, x] \neq 0$:
We are in the setting of Figure 4.3 (left). Since $\dim(\langle x, d\rangle) = 2$ there is $y \in \langle x, d\rangle \cap b^\perp = \langle c, d\rangle \cap b^\perp$. Furthermore, as $a^\perp \cap \langle c, d\rangle = \langle x\rangle$, it holds that $[a, y] \neq 0$. Then for $\Omega' = \{a, b, x, y\}$ we have $O(\Omega) = O(\Omega')$ and the graph $G(\Omega')$ is a 4-cycle (see Figure 4.3 (right)). □

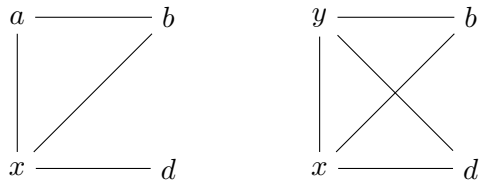


Figure 4.2: Orthogonality graphs occurring in case (1) of Lemma 4.3.9.

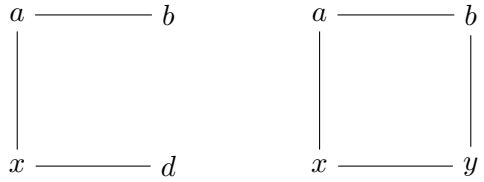


Figure 4.3: Orthogonality graphs occurring in case (2) of Lemma 4.3.9.

Lemma 4.3.10 (Non-existence of a Mermin-square in the qudit world)**.** *If* $\Omega = \{a, b, x, y\}$ *with* $\dim(\{a, b, x, y\}) = 4$ *and* $G(\Omega)$ *is a* 4*-cycle, then* $O(\Omega) = \langle a, b, x, y\rangle$.

Proof, based on Lemma 1 in [DOBV+17]. The orthogonality relations are depicted in Figure 4.3 (right). Since $\langle a\rangle, \langle b\rangle, \langle x\rangle, \langle y\rangle \subset O(\Omega)$, we may assume that $[a, y] = [b, x] = 1$. It suffices to prove that the planes spanned by non-orthogonal elements are contained in $O(\Omega)$, i.e.

$$\langle a, y\rangle, \langle b, x\rangle \subset O(\Omega). \tag{4.14}$$

Then each point $x \in \langle a, b, x, y\rangle$ can be written as

$$x = \underbrace{(\alpha a + \delta y)}_{\in O(\Omega)} + \underbrace{(\beta b + \gamma x)}_{\in O(\Omega)}$$

and, due to $[a, b] = [a, x] = [d, y] = [x, y] = 0$, it holds that

$$[\alpha a + \delta y, \beta b + \gamma x] = 0.$$

Now (4.14) follows immediately from the arguments in the proof of Lemma 1 in [DOBV$^+$17]. □

We will continue with the proof of Lemma 4.3.7.

Proof of Lemma 4.3.7. Let $\Omega \subset \mathbb{F}_d^{2n}$ be a subspace. If $v \in \Omega$, there is nothing to show, so assume $v \notin \Omega$. Clearly,

$$\langle v, \Omega \cap v^\perp \rangle \subseteq O(\Omega \cup \{v\}) \tag{4.15}$$

and if $\Omega \subset v^\perp$ we have equality in (4.15) and $O(\Omega \cup \{v\})$ is a subspace.

So assume that there is $x \in \Omega$ such that $[v, x] \neq 0$. Then $\dim(\Omega \cap v^\perp) = \dim(\Omega) - 1$ and it follows that $\Omega = \langle x, \Omega \cap v^\perp \rangle$. Now we consider two cases:

(1) If $\Omega \cap v^\perp$ is isotropic, then also $\langle v, \Omega \cap v^\perp \rangle$. Further, $H = \langle v, \Omega \cap v^\perp \rangle \cap x^\perp$ is isotropic with $v, x \notin H$ and $\dim(H) = \dim(\Omega) - 1$. We obtain

$$O(\Omega) = \langle x, H \rangle \cup \langle v, H \rangle,$$

so $O(\Omega)$ is of the form (4.13).

(2) If $\Omega \cap v^\perp$ is not an isotropic subspace, then there are $a, b \in \Omega \cap v^\perp$ with $[a, b] \neq 0$.

(2.1) If $a, b \in x^\perp$, then $G(\{a, b, v, x\})$ is a 4-cycle; see Figure 4.4 (left) and therefore $O(\{a, b, v, x\}) = \langle a, b, v, x \rangle$, by Lemma 4.3.10.

(2.2) If $\{a, b\} \not\subseteq x^\perp$, we may assume without loss of generality that $[a, x] = 0$ and $[b, x] \neq 0$, due to $\dim(\langle a, b \rangle) = 2$. Now we are in the setting of Figure 4.4 (right) and precisely in case (2) of the proof of Lemma 4.3.9. Again using Lemma 4.3.10, we obtain $O(\{a, b, v, x\}) = \langle a, b, v, x \rangle$.

Consequently, $\langle v, x \rangle \subset O(\Omega \cup \{v\})$ for both cases (2.1) and (2.2). As x was chosen arbitrarily in $\Omega \setminus v^\perp$, it follows $O(\Omega \cup \{v\}) = \langle \Omega, v \rangle$. □

To finally prove Lemma 4.3.8, we need one more observation, distilled from the proof of Lemma 4.3.7.

Corollary 4.3.11. *If $\Omega \subset \mathbb{F}_d^{2n}$ is a subspace and contains $a, b, c, d \in \Omega$ such that the graph $G(\{a, b, c, d\})$ is a 4-cycle and $\dim(\langle a, b, c, d \rangle) = 4$, then $O(\Omega \cup \{v\}) = \langle \Omega, v \rangle$ for all $v \in \mathbb{F}_d^{2n}$.*

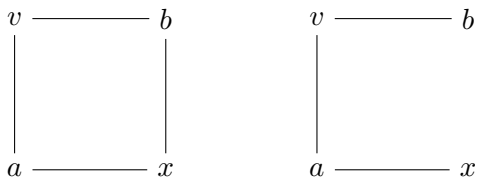


Figure 4.4: Orthogonality graphs occurring in case (2) of Lemma 4.3.7.

Proof. It suffices to prove that $\Omega \cap v^\perp$ is not isotropic because then case (2) of the proof of Lemma 4.3.7 can be applied. Therefore, let $I \subset \Omega$ be an isotropic subspace. As the largest isotropic subspace contained in $\langle a, b, c, d\rangle$ has dimension 2, it follows that $\dim(\langle a, b, c, d\rangle \cap I) \leq 2$. Since $\langle a, b, c, d\rangle$ is a 4-dimensional subspace contained in Ω and $I \subset \Omega$, it follows $\dim(I) \leq \dim(\Omega) - 2$. However, $\Omega \cap v^\perp$ is a $(\dim(\Omega) - 1)$-dimensional subspace, hence, $\Omega \cap v^\perp$ is not isotropic. □

Finally, we will complete the proof of Proposition 4.3.6 by proving Lemma 4.3.8.

Proof of Lemma 4.3.8. Let

$$\Omega = \langle I, h_1\rangle \cup \cdots \cup \langle I, h_\ell\rangle,$$

where I is an isotropic subspace, $h_i \in I^\perp$ and $[h_i, h_j] \neq 0$ for $i \neq j$. We assume that $\ell > 1$, otherwise Ω is a subspace and we are in the situation of Lemma 4.3.7. As before, there is nothing to show if $v \in \Omega$, so let $v \notin \Omega$.

We will do a case distinction:

(1) $I \subset v^\perp$:

(1.1) If $[v, h_i] \neq 0$ for all $i = 1, \ldots, \ell$, then

$$O(\Omega \cup \{v\}) = \langle I, h_1\rangle \cup \cdots \cup \langle I, h_\ell\rangle \cup \langle I, v\rangle.$$

(1.2) Otherwise, we may assume that $[h_1, v] = 0$. Further, we may assume that $[h_2, v] = 0$; if this was not the case, we could replace v by $\tilde{v} \in \langle h_1, v\rangle^\times \subset O(\Omega)$ such that $[\tilde{v}, h_2] = 0$.

(1.2.1) If $[h_2, v] = \cdots = [h_\ell, v] = 0$, then $\tilde{I} := \langle v, I\rangle$ is an isotropic subspace and

$$O(\Omega \cup \{v\}) = \langle \tilde{I}, h_1\rangle \cup \cdots \cup \langle \tilde{I}, h_\ell\rangle.$$

(1.2.2) If $[v, h_j] \neq 0$ for some $j \in \{3, \dots, \ell\}$, then there is $v' \in \langle h_1, v\rangle^\times$ such that $[v', h_j] = 0$, without loss of generality $j = 3$ and $[v, h_3] \neq 0$. Since $[v, h_2] = 0$ and $v' = \alpha h_1 + \beta v \in \langle h_1, v\rangle$ with $\alpha, \beta \in \mathbb{F}_d^\times$, it follows that

$$[v', h_2] = \alpha[h_1, h_2] \neq 0.$$

Since v' is contained in the isotropic subspace $\langle h_1, v\rangle$ the orthogonality graph $G(\{v, v', h_2, h_3\})$ is given in Fig. 4.5.

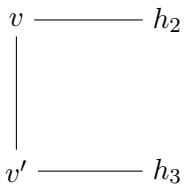


Figure 4.5: Orthogonality graph occurring in case (1.2.2) of Lemma 4.3.8.

Applying case (2) of Lemma 4.3.9 and Lemma 4.3.10 shows that

$$\langle v, v', h_2, h_3\rangle = O(\{v, v', h_2, h_3\}) \subset O(\Omega \cup \{v\})$$

and therefore, due to $v, v', h_2, h_3 \in I^\perp$, the subspace $U' = \langle I, \tilde{v}, v', h_2, h_3\rangle$ is contained in $O(\Omega)$. To conclude, observe that

$$O(\Omega) = O(U' \cup \{h_1\} \cup \{h_4\} \cup \cdots \cup \{h_\ell\}).$$

Now consider $O(U' \cup \{h_1\})$ and use the fact that we can construct a 4-cycle in $\langle v, v', h_2, h_3\rangle \subset U'$. Thus, we are able to apply Corollary 4.3.11 to get $O(U' \cup \{h_1\}) = \langle U', h_1\rangle$ and iteratively

$$O(\Omega) = O(U' \cup \{h_1\} \cup \{h_4\} \cdots \cup \{h_\ell\}) = \langle I, h_1, \dots, h_\ell, v\rangle.$$

(2) $I \not\subseteq v^\perp$:
Let $u \in I$ with $[u, v] \neq 0$. Since $h_1, h_2 \in I^\perp \subset u^\perp$ we may choose $\tilde{h}_1 \in \langle u, h_1\rangle^\times$ and $\tilde{h}_2 \in \langle u, h_2\rangle^\times$ such that $[\tilde{h}_1, v] = [\tilde{h}_2, v] = 0$. Since $[\tilde{h}_1, \tilde{h}_2] = [h_1, h_2] \neq 0$, the orthogonality graph $G(u, v, \tilde{h}_1, \tilde{h}_2)$ is given in Fig. 4.6 and therefore a 4-cycle. Hence, by Lemma 4.3.10, it follows that $O(\Omega \cup \{v\})$ contains the subspace $\langle \tilde{h}_1, \tilde{h}_2, u, v\rangle$. Now we can iteratively add the remaining elements of I and the cosets $\langle h_3.I\rangle, \dots \langle h_\ell, I\rangle$ to $\langle \tilde{h}_1, \tilde{h}_2, u, v\rangle$ and apply Corollary 4.3.11 to obtain $O(\Omega) = \langle I, h_1, \dots, h_\ell, v\rangle$. □

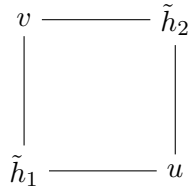


Figure 4.6: Orthogonality graph occurring in case (2) of Lemma 4.3.8.

4.3.4 Proof of Theorem 4.3.2

Finally, we will put all pieces together to prove the main result of this chapter. First, we will show that if A_Ω^η is as in Theorem 4.3.2, then it is contained in Λ_n. To see this, consider an isotropic subspace I and $\gamma \in I^*$. Then, by character orthogonality,

$$\mathrm{Tr}(A_\Omega^\eta \Pi_I^\gamma) = \frac{1}{d^n} \sum_{u \in \Omega \cap I} \omega^{\eta(u)} \omega^{-\gamma(u)} = \delta_{\eta_{|\Omega\cap I} = \gamma_{|\Omega \cap I}} \geq 0$$

because $\Omega \cap I$ is an isotropic subspace and $\eta_{|\Omega\cap I} \in (\Omega \cap I)^*$, as a consequence of Theorem 4.3.1. This holds in particular for the case where Π_I^γ is a stabilizer state, hence, $A_\Omega^\eta \in \Lambda_n$.

Now assume that Ω is a subspace or of the form (4.13) and $A_\Omega^\eta \in \Lambda_n$, so $\eta : \Omega \to \mathbb{F}_d$ and $\eta_{|I} \in I^*$ for all isotropic subspaces I contained in Ω. To characterize all possible η, we will again use Lemma 1 of [DOBV$^+$17]. It says that if Ω is a subspace with $\dim(\Omega) \geq 2$ and $\eta : \Omega \to \mathbb{F}_d$ such that $\eta(a+b) = \eta(a) + \eta(b)$ whenever $[a,b] = 0$, then η is linear on Ω, i.e. $\eta \in \Omega^*$. This directly shows (i) of Theorem 4.3.2.

Similarly, for (ii), if Ω is of the form (4.13) and $a_1, \ldots, a_k$ is a basis of I, then η is uniquely determined by $\eta(a_1), \ldots, \eta(a_k), \eta(h_1), \ldots, \eta(h_\ell) \in \mathbb{F}_d$.

4.3.5 Classification of maximal sets of mutually non-orthogonal elements

Sets $\Omega \subset \mathbb{F}_d^{2n}$ of the form (4.13) come with a set of mutually non-orthogonal elements $h_1, \ldots, h_\ell \in \mathbb{F}_d^{2n}$. This raises the question about the maximal number of mutually non-orthogonal elements in $\mathbb{F}_d^{2n}$. This number will turn out to be $dn+1$. The proof is as a generalization of the corresponding qubit construction in [RBVT$^+$20, Theorem 1].

Let $e_1, \dots, f_1, \dots, f_n \in \mathbb{F}_d^{2n}$ be a basis of the symplectic vector space $\mathbb{F}_d^{2n}$ with $[e_i, e_j] = [f_i, f_j] = 0$ and $[e_i, f_j] = \delta_{i,j}$. For example we may choose $e_i(k) = \delta_{ik}$ and $f_j(k) = \delta_{j+n,k}$ for $i, j = 1, \dots, n$ and $k = 1, \dots, 2n$.

First, observe that the 2-dimensional space $\mathbb{F}_d^2$ can be decomposed into $d+1$ mutually non-orthogonal lines:

$$\mathbb{F}_d^2 = \langle a_1 \rangle \cup \cdots \cup \langle a_{d+1} \rangle, \quad [a_i, a_j] \neq 0 \text{ for } i \neq j. \tag{4.16}$$

For example, such a decomposition if given by

$$\mathbb{F}_d^2 = \langle e_1 \rangle \cup f_1 \rangle \cup \langle e_1 + f_1 \rangle \cup e_1 + 2f_1 \cup \cdots \cup \langle e_1 + (d-1)f_1 \rangle.$$

This gives rise to the following construction for general n (which is a generalization of the construction given in [RBVT+20, Equations. (16), (17)]): Label the elements of $\mathbb{F}_d^{2n}$ by $(u_1, \dots, u_n)$ with $u_i \in \mathbb{F}_d^2$ for $i = 1, \dots, n$.

Lemma 4.3.12. *Every maximal set of mutually non-orthogonal elements in $\mathbb{F}_d^{2n}$ can be mapped to the following $dn+1$ elements by a symplectic transformation:*

$$\begin{aligned}
&(a_1, 0, \dots, 0), (a_2, 0, \dots, 0), \dots, (a_d, 0, \dots, 0),\\
&(a_{d+1}, a_1, 0 \dots, 0), (a_{d+1}, a_2, 0, \dots, 0), \dots, (a_{d+1}, a_d, 0, \dots, 0),\\
&(a_{d+1}, a_{d+1}, a_1, 0 \dots, 0), (a_{d+1}, a_{d+1}, a_2, 0, \dots, 0), \dots, (a_{d+1}, a_{d+1}, a_d, 0, \dots, 0),\\
&\vdots\\
&(a_{d+1}, a_{d+1} \dots, a_{d+1}, a_1), (a_{d+1}, a_{d+1} \dots, a_{d+1}, a_2), \dots, (a_{d+1}, a_{d+1} \dots, a_{d+1}, a_d),\\
&(a_{d+1}, a_{d+1} \dots, a_{d+1}, a_{d+1}),
\end{aligned}$$

where a_i is as in (4.16).

Proof. It is straightforward to check that all given generators are mutually non-orthogonal: The symplectic inner product of two elements in the same line is always of the form $[a_i, a_j] \neq 0$ for $1 \leq i \neq j \leq d$; if two elements lie in distinct lines, then it is of the form $[a_i, a_{d+1}] \neq 0$ for $i = 1, \dots, d$.

Now let $g_1, \dots, g_k$ be any set of k mutually non-orthogonal elements with $k \leq dn+1$. By Witts theorem [Asc00, Section 20], the elements

$g_1, \ldots, g_k$ can be mapped to a subset of the elements given in the lemma. Thus, it suffices to show that there is no element that is mutually non-orthogonal to all $dn + 1$ elements of the lemma.

Let $u = (u_1, \ldots, u_n)$. Due to (4.16), for every u_i, there is $j_i \in \{1, \ldots, d+1\}$ such that $[u_i, a_{j_i}] = 0$. Hence, to have non-zero symplectic inner product with the first k lines of the set of elements in the lemma, it must hold that $u_1 = \cdots = u_k = a_{d+1}$. However, for $k = n$ this implies that $u = (a_{d+1}, \ldots, a_{d+1})$, which is precisely the last element. □

4.4 Classical simulation

Finally, we will argue why one should consider all operators $A^\eta_\Omega \in \Lambda_n$, which satisfy the conditions of Theorem 4.3.2, as classical objects. This is due to the fact that each such A^η_Ω has an efficient classical description using a set of generators of Ω. If Ω is a subspace, then A^η_Ω is fully described by a basis $a_1, \ldots, a_k$ of Ω and the images $\eta(a_1), \ldots, \eta(a_k) \in \mathbb{F}_d$. For Ω being of the form (4.13), i.e.

$$\Omega = \langle I, h_1 \rangle \cup \cdots \cup \langle I, h_\ell \rangle$$

we can fully characterize A^η_Ω by $h_1, \ldots, h_\ell$, a basis $a_1, \ldots, a_k$ of the isotropic subspace I and the images $\eta(h_i), \eta(a_j) \in \mathbb{F}_d$. As a consequence of Lemma 4.3.5, the maximal number of non-orthogonal elements is $dn + 1$, so again there is a classical linear description of A^η_Ω.

Next, we will shortly sketch how to classically update the operators A^η_Ω under Clifford unitaries and Pauli measurements. The ideas are based on References [SA04, VFGE12, RBVT+20].

Clifford unitaries. If d is an odd prime, then every Clifford unitary can be uniquely described by some $a \in \mathbb{F}_d^{2n}$ and a symplectic linear map $S : \mathbb{F}_d^{2n} \to \mathbb{F}_d^{2n}$, i.e. a linear map that preserves the symplectic inner product $[\cdot, \cdot]$. The action of a Clifford unitary $U_{S,a}$ on a generalized Pauli matrix is given by [Gro06, Theorem 3]

$$U_{S,a} w(u) U_{S,a}^\dagger = \omega^{[a,u]} w(Su). \tag{4.17}$$

Hence, to get a classical description of $U_{F,a}A^{\eta}_{\Omega}U_{F,a}$ we only need to update the generators of Ω and their images correctly. For instance, if Ω is of the form (4.13), with $I = \langle b_1, \ldots, b_k\rangle$ then

$$U_{F,a}A^{\eta}_{\Omega}U^{\dagger}_{F,a} = A^{\eta'}_{\Omega'}$$

where

$$\Omega' = \langle S(I), S(h_1)\rangle \cup \cdots \cup \langle S(I), S(h_\ell)\rangle$$

and

$$\eta'(S(b_i)) = \eta(b_i) + [a, b_i], \quad \eta'(S(h_i)) = \eta(h_i) + [a, h_i].$$

In the same fashion, we can update A^{η}_{Ω} when Ω is a subspace.

Pauli measurements. To classically compute the update of A^{η}_{Ω} under a Pauli measurement, consider some Pauli observable $w(a)$. The corresponding measurement projectors are given by (see Equation (4.7))

$$\{d^{n-1}\Pi^{\gamma}_{\langle a\rangle} \,:\, \gamma \in \langle a\rangle^*\}.$$

We use the ideas from [RBVT$^+$20, Lemma 5] and [ZORH21, Lemma 7] to describe the effect of measuring the observable $w(a)$ on A^{η}_{Ω}. we distinguish two cases:

(1) If $a \in \Omega$, then $\mathrm{Tr}(d^{n-1}\Pi^{\gamma}_{\langle a\rangle}A^{\eta}_{\Omega}) = \delta_{\eta(a)=\gamma(a)}$, so the post-measurement operator will be deterministically

$$\Big(d^{n-1}\Pi^{\eta_{|\langle a\rangle}}_{\langle a\rangle}\Big)A^{\eta}_{\Omega}\Big(d^{n-1}\Pi^{\eta_{|\langle a\rangle}}_{\langle a\rangle}\Big) = \frac{1}{d^n}\sum_{u\in\Omega\cap a^{\perp}}\omega^{\eta(u)}w(u) = A^{\eta_{|\Omega\cap a^{\perp}}}_{\Omega\cap a^{\perp}}. \quad (4.18)$$

The set $\Omega \cap a^{\perp}$ is again either a subspace or of the form (4.13) and can be computed efficiently.

(2) If $a \notin \Omega$, then $\mathrm{Tr}(d^{n-1}\Pi^{\gamma}_{\langle a\rangle}A^{\eta}_{\Omega}) = 1/d$ for all $\gamma \in \langle a\rangle^*$. Then for Ω of the form (4.13), the post-measurement operator becomes with probability $1/d$ for all d linear functions $\gamma \in \langle a\rangle^*$:

$$\frac{\Big(d^{n-1}\Pi^{\gamma}_{\langle a\rangle}\Big)A^{\eta}_{\Omega}\Big(d^{n-1}\Pi^{\gamma}_{\langle a\rangle}\Big)}{\mathrm{Tr}(d^{n-1}\Pi^{\gamma}_{\langle a\rangle}A^{\eta}_{\Omega})} = \frac{1}{d^n}\sum_{i=1}^{\ell}\sum_{u\in\langle a,h_i,I\rangle\cap a^{\perp}}\omega^{\gamma*\eta(u)}w(u) = A^{\gamma*\eta}_{\Omega'}, \quad (4.19)$$

where

$$\Omega' = (\langle a, h_1, I\rangle \cap a^\perp) \cup \cdots \cup (\langle a, h_\ell, I\rangle \cap a^\perp)$$

is again of the form (4.13). The map $\gamma * \eta$ is given by

$$\gamma * \eta : \Omega' \to \mathbb{F}_d, \quad \gamma * \eta(ka + u) = \gamma(ka) + \eta(u), \quad k \in \mathbb{F}_d,\, u \in \Omega \cap a^\perp. \tag{4.20}$$

Both Ω' and $\gamma * \eta$ can be efficiently classically computed. In the same fashion, if Ω is a subspace, then

$$\frac{\left(d^{n-1}\Pi^\gamma_{\langle a\rangle}\right) A^\eta_\Omega \left(d^{n-1}\Pi^\gamma_{\langle a\rangle}\right)}{\mathrm{Tr}(d^{n-1}\Pi^\gamma_{\langle a\rangle} A^\eta_\Omega)} = \sum_{u\langle a, \Omega\cap a^\perp\rangle} \omega^{\gamma*\eta(u)} w(u) = A^{\gamma*\eta}_{\Omega'} \tag{4.21}$$

with subspace $\Omega' = \langle a, \Omega \cap a^\perp\rangle$ and $\gamma * \eta$ as in (4.20).

In summary, to simulate a measurement of $w(a)$ on A^η_Ω with $a \in \Omega$, we update A^η_Ω according to Equation (4.18). If $a \notin \Omega$, we pick $\gamma \in \langle a\rangle^*$ uniformly at random and update A^η_Ω according to Equations (4.19) and (4.21).

Classical simulation of QCM. Finally, we describe how to use the update rules to classically simulate QCM [3] for an input spate ρ which is contained in

$$P := \mathrm{conv}\{A^\eta_\Omega \in \Lambda_n \,:\, A^\eta_\Omega = \sum_{u\in\Omega} \omega^\eta(u) w(u), \eta : \Omega \to \mathbb{F}_d^{2n}\}. \tag{4.22}$$

The set P is a polytope, as there are only finitely many tuples (Ω, η). So,

$$\rho = \sum_{\Omega,\eta} \lambda_{\Omega,\eta} A^\eta_\Omega, \quad \lambda_{\Omega,\eta} \geq 0, \quad \sum_{\Omega,\eta} \lambda_{\Omega,\eta} = 1,$$

and we can classically simulate the evolution of ρ under Clifford unitaries and Pauli measurements in the following way: We sample A^η_Ω from the probability distribution $\{\lambda_{\Omega,\eta}\}$ and then compute the evolution of the sampled A^η_Ω under Clifford unitaries and Pauli measurements. The sketched algorithm correctly reproduces the outcomes

[3] This amounts to updating ρ under Clifford unitaries and Pauli measurements.

of the corresponding quantum procedure; for a proof, see for example [RBVT$^+$20, Theorem 3] or [ZORH21, Theorem 2] [4]. This algorithm is runs in polynomial time in the number of qubits, provided that we can sample from the distribution $\{\lambda_{\Omega,\eta}\}$ in polynomial time in the number of qubits.

By far, the polytope P does not contain all quantum states, but it strictly contains the Wigner simplex

$$\mathrm{conv}\{A^{\eta}_{\mathbb{F}_d^{2n}} \,:\, \eta \in (F_d^{2n})^*\}.$$

This description of P in (4.22) is over-complete, i.e. not all A^{η}_{Ω} are vertices of P. In fact, it suffices to consider *inclusion maximal* sets. That is, if $A^{\eta}_{\Omega} \in \Lambda_n$ and Ω is of the form (4.13), then

$$A^{\eta}_{\Omega} \in \mathrm{conv}\{A^{\eta'}_{\Omega'} \in \Lambda_n \,:\, \Omega' \text{ of the form (4.13)},\ \Omega \subset \Omega', \eta'_{|\Omega} = \eta\}.$$

In the same fashion, if $A^{\eta}_{\Omega} \in \Lambda_n$ and Ω is a subspace, then

$$A^{\eta}_{\Omega} \in \mathrm{conv}\{A^{\eta'}_{\Omega'} \in \Lambda_n \,:\, \Omega \text{ subspace},\ \Omega \subset \Omega', \eta'_{|\Omega} = \eta\}.$$

For proofs we refer the reader to Lemma 1 in [RBVT$^+$20] and Lemma 5 in [ZORH21] [5].

4.5 Conclusion

In this chapter, we have characterized a particular class of operators that live in the Λ-polytope. Furthermore, we have argued why these operators should be considered as classical objects. This is due to the fact that they allow an efficient description in terms of their generators. Additionally, updating these operators under Clifford unitaries and Pauli measurements can be efficiently classically simulated.

If the goal is to identify or classify non-classical, respectively quantum structures in the Λ-polytope, our findings can be seen as a rather negative result – the considered family of operators do not capture this. This aligns with the results for qubit systems, despite the slightly

[4] The references deal with the qubit case; however this can be straightforwardly adapted to qudits.

[5] Lemma 1 in [RBVT$^+$20] is only for qubits and Lemma 5 in [ZORH21] only for isotropic subspaces, but both cases can be adapted straightforwardly.

different structure of the corresponding Λ-polytopes for qubits and qudits.

Hence, the question remains open: Is it possible to characterize vertices of Λ or other interesting elements in Λ, which should be considered non-classical?

Part II

The Lattice World

Chapter 5

Introduction - Lattices

An n-dimensional lattice $L \subset \mathbb{R}^n$ is a discrete subgroup of $\mathbb{R}^n$ which spans $\mathbb{R}^n$, or equivalently, the set of all integer linear combinations of a set of n linearly independent vectors $v_1, \ldots, v_n \in \mathbb{R}^n$.

Lattices and more generally point configurations in Euclidean space have been a central object of mathematical research over the last centuries. Arising from early work of Lagrange, Gauss, Hermite, Korkin and Zolotarev on the reduction theory of quadratic forms, their study has proven fruitful in various areas of mathematics. Recently, Maryna Viazovska was awarded the fields medal for, among other major contributions, showing that the densest sphere packings in dimensions 8 and 24 are induced by the lattices E_8 and Λ_{24} [Via17, CKM+16].

In this thesis, we will encounter two types of optimization problems related to lattices. The first one deals with optimization in the space of lattices. That is, the space of lattices is our search space and our goal is to find a lattice which is optimal for our optimization problem. In contrast, in the second case, we are given a lattice and we aim to determine a certain invariant of this lattice. This invariant can be expressed as an optimization problem.

Chapter 6: Optimization in the space of lattices

Our setup is as follows. Let $f : (0, \infty) \to \mathbb{R}$ be some function and let L be a lattice, then the *f-potential energy* of L is defined as

$$\mathcal{E}(f, L) = \sum_{x \in L \setminus \{0\}} f(\|x\|^2). \tag{5.1}$$

One can think of L describing a system of particles and the energy between each tuple $x, y \in L$ is determined by $f(||u||^2)$, where $u = x - y \in L$. For the optimization problem we want to consider, we fix some function f and we are interested in a lattice L that minimizes/maximizes $\mathcal{E}(f, L)$. Phrasing the task in this way is a priori not well-defined for two reasons.

First, it might occur that the right hand side does not converge for the choice of f. We will focus on Gaussian potential functions

$$f_\alpha(r) = e^{-\alpha r}, \quad \alpha > 0.$$

In this case, the right hand side of (5.1) converges for every lattice L and every $\alpha > 0$. Gaussian potential functions appear naturally in the study of lattices. For example, they can be used to design algorithms for solving the shortest respectively closest vector problem [ADSD15, ADRSD15] or to give bounds on lattice parameters [RSD17]. When particles interact according to a Gaussian potential function, this is referred to as the Gaussian core model [Sti].

The second obstruction is that we can make the f-potential energy $\mathcal{E}(f_\alpha, L)$ arbitrarily large/small by simply scaling the lattice. To deal with this, we restrict ourselves to lattices that are "comparable". This means the following: Let $\mathcal{B}(L) = \{v_1, \ldots, v_n\}$ be a basis of L. Then the (symmetric positive semidefinite) $n \times n$ matrix

$$M_L(\mathcal{B})_{i,j} = v_i^\mathsf{T} v_j$$

is called a *Gram matrix* of L. The Gram matrix is dependent on the choice of the basis, however, the *determinant* of the lattice, $\det(L) := \det(M_L(\mathcal{B}))$ is a lattice invariant independent of the choice of basis. Furthermore, it is invariant under orthogonal transformations of the

lattice, that is $\det(AL) = \det(L)$ for all orthogonal matrices A. To make lattices comparable, we only consider lattices L with $\det(L) = 1$. In dimension 8 and 24, Cohn et al [CKM$^+$22] show that the root lattice E_8 and the Leech lattice Λ_{24} are global minimizers for Gaussian potential functions among all lattices of determinant one. In fact, their result is even stronger. They show that E_8 and Λ_{24} minimize the f-potential energy among all point configurations of point density 1 and for all completely monotonic functions; see [CKM$^+$22, Section 1] for definitions of point density and complete monotonicity. Such global minimizers are referred to as *ground states* for the given function f.

Moreover, restricting the search space to Gram matrices of lattices is no loss of generality when considering $\mathcal{E}(f, L)$. The f-potential is obviously invariant under the action of the orthogonal group on L, i.e. $\mathcal{E}(f, AL) = \mathcal{E}(f, L)$ for every orthogonal matrix A. This allows us to see the potential energy as a function $L \mapsto \mathcal{E}(f, L)$ where we parametrize lattices by the manifold of positive definite matrices with determinant one.

In Chapter 6 we will use this to conduct a local analysis of $\mathcal{E}(f, L)$. Parameterizing the space of lattices with point density one as a manifold makes it possible to compute the gradient and the Hessian of $\mathcal{E}(f, L)$ [Cou06, CS12].

A lattice is critical for $\mathcal{E}(f, L)$ whenever the gradient vanishes. A simple sufficient condition for a vanishing gradient is that all non-empty subsets

$$L(r) = \{x \in L \,:\, x^\mathsf{T} x = r\}$$

of the lattice form *spherical 2-designs*. If $L(r) \neq \emptyset$, then $L(r)$ is called a shell of L. A finite point set $X \subset \{x \in \mathbb{R}^n \,:\, x^\mathsf{T} x = r\}$ is a spherical t-design if it satisfies the cubature rule

$$\frac{1}{|X|} \sum_{x \in X} f(x) = \int_{x^\mathsf{T} x = r} f(x) \mathrm{d}\sigma(x)$$

for all polynomials $f : \mathbb{R}^n \to \mathbb{R}$ of degree at most t. The corresponding measure $\sigma(x)$ is the Haar-measure on the sphere of radius r.

and have been extensively studied in the last few decades (for a survey see [BB09]).

There is a particularly well-behaved class of lattices where the above property oftentimes holds, namely *even unimodular lattices*. These lattices satisfy $x^\mathsf{T} x \in 2\mathbb{Z}$ for all $x \in L$ (even) and $L = L^*$ (unimodular) where

$$L^* = \{y \in \mathbb{R}^n : x^\mathsf{T} y \in \mathbb{Z} \text{ for all } x \in L\}$$

is the *dual lattice* of L. Even unimodular lattices only exist in dimension divisible by 8. If their dimension is $8, 16$ or 24, then the shells automatically form spherical 2-designs [Ebe94, Chapter 3].

In the case of even unimodular lattices we can use the theory of modular forms to analyze the behavior of the gradient and Hessian of $\mathcal{E}(f, L)$. Modular forms are functions that satisfy particular transformation properties and appear naturally in number theory and complex analysis.

To illustrate how to make use of modular forms, rewrite the Gaussian potential energy $\mathcal{E}(f_\alpha, L)$ as a sum over the shells, that is

$$\mathcal{E}(f_\alpha, L) = \sum_{m=0}^{\infty} a_m e^{-\alpha m} \quad \text{with} \quad a_m = |L(2m)|.$$

Then $\mathcal{E}(f_\alpha, L) = \Theta_L(\alpha i/\pi) - 1$, where the function

$$\Theta_L : \{\tau \in \mathbb{C} : \operatorname{Im}(z) > 0\} \to \mathbb{C}, \quad \Theta_L(\tau) = \sum_{m=0}^{\infty} a_m q^m, \quad q = e^{2\pi i \tau} \tag{5.2}$$

As a consequence of the Poisson summation formula [Ebe94, Chapter 2], one can relate the theta function of an even unimodular lattice to its dual lattice:

$$\Theta_L(iy) = y^{-n/2} \Theta_{L^*}(i/y) \quad \text{for} y > 0. \tag{5.3}$$

Additionally, it is easy to check that

$$\Theta_L(iy + b) = \Theta_L(iy) \quad \text{for all } b \in \mathbb{Z}. \tag{5.4}$$

Holomorphic functions $f : \{z \in \mathbb{C} : \operatorname{Im}(z) > 0\} \to \mathbb{C}$ that have a power series expansion of the form (5.2) and satisfy (5.3) and (5.4) are called *modular forms of weight* $n/2$. For comparably small n [1],

[1] There is an explicit dimension formula for the vector space of modular forms of weight k, see for example [KK07, Page 151].

the vector space of modular forms of weight $n/2$ is low dimensional and the occurring coefficients $a_m = |L(2m)|$ are well studied [JR11]. Using the structural properties of the vector spaces of modular forms, we derive a formula for the eigenvalues of the Hessian of $\mathcal{E}(f_\alpha, L)$ when L is an even unimodular lattice of dimension $n \geq 32$. When L is an even unimodular lattice and all shells are 4-designs, the Hessian is a scalar multiple of the identity and its eigenvalues can be expressed as a formula that only depends on α, n and a_m, i.e. the size of the shells $|L(2m)|$ for all $r > 0$.

If the shells are only 2-designs, we can still compute the eigenvalues of the Hessian but this requires substantially more work. Up to isomorphism, all even unimodular lattices in dimension $n \leq 24$ are uniquely determined by their root sublattices. The root sublattice is the sublattice of L that is spanned by the *roots* $L(2)$ where $2 = \lambda(L)$ is the length of the shortest vector in L. The reflections $I - 2xx^\mathsf{T}$ define a group action on $L(2)$. This group is called the *Weyl group* of the root system $L(2)$. Now the eigenvalues of the Hessian can be computed with the help of elementary representation theory for the corresponding Weyl group.

In the case of even unimodular lattices in dimension 32, where all shells are 4-designs, and the Hessian of $\mathcal{E}(f_\alpha, L)$ is a scalar multiple of the identity, we show that the only eigenvalue of the Hessian becomes negative for the parameter $\alpha = \pi$. This implies that there are even unimodular lattices that are *local maximizers* for $\mathcal{E}(f_\alpha, L)$.

Our result has the following consequence. In lattice theory, it is an open question whether

$$\mathcal{E}(f_\alpha, L) \leq \mathcal{E}(f_\alpha, \mathbb{Z}^n) \tag{5.5}$$

holds for all α and all *stable* lattices L, that is $\det(L) = 1$ and $\det(L') \geq 1$ for all sublattices $L' \subseteq L$ of L [RSD17, ERS22]. Statements of this flavor have been shown for other functions [ERS22] and fixed parameters α [RSD17]. The underlying proof technique is to show that maximizers for the Gaussian potential energy in the (compact) set of stable lattices lie on the boundary of this set. Then an inductive argument is applied to show that these lattices are isomor-

phic to $\mathbb{Z}^n$. However, the existence of local maxima for the f_α-potential energy shows that such a proof technique cannot be directly applied to prove (5.5).

Chapter 7: Least distortion embeddings of flat tori

For an n-dimensional lattice L, the quotient space $\mathbb{R}^n/L$ is a called a *flat torus*. The quotient $\mathbb{R}^n/L$ is equipped with the following metric:

$$d_{\mathbb{R}^n/L}(x,y) = \min_{v\in L} |x-y-v|,$$

where $|\cdot|$ is the standard norm on $\mathbb{R}^n$. An embedding of $\mathbb{R}^n/L$ is an injective continuous map $\varphi : \mathbb{R}^n/L \to H$ to a Hilbert space H. We are interested in an embedding $\varphi : \mathbb{R}^n/L \to H$ such that the metric on H approximates the metric $d_{\mathbb{R}^n/L}$ as good as possible. More formally, we want to find a Hilbert space H and an injective function $\varphi : \mathbb{R}^n/L \to H$ such that the *distortion*

$$\operatorname{dist}(\varphi) = \sup_{\substack{x,y\in\mathbb{R}^n/L\\ x\neq y}} \frac{\|\varphi(x)-\varphi(y)\|}{d_{\mathbb{R}^n/L}(x,y)} \cdot \sup_{\substack{x,y\in\mathbb{R}^n/L\\ x\neq y}} \frac{d_{\mathbb{R}^n/L}(x,y)}{\|\varphi(x)-\varphi(y)\|},$$

where $||\cdot||$ denotes the metric on the Hilbert space, becomes as small as possible. The first factor of the above product is called the *expansion* of φ and the second factor the *contraction of* φ. In the language of optimization our goal is to compute

$$c_2(\mathbb{R}^n/L) := \inf\{\operatorname{dist}(\varphi) \ : \ \varphi\colon \mathbb{R}^n/L \to H \text{ for some Hilbert space } H,\ \varphi \text{ injective}\}.$$

The simplest example of a flat torus is $\mathbb{R}/\mathbb{Z}$. To construct an embedding of $\mathbb{R}/\mathbb{Z}$, we choose the interval $[-\frac{1}{2},\frac{1}{2}]$ with $-\frac{1}{2} = \frac{1}{2} \pmod{\mathbb{Z}}$ as a fundamental domain for $\mathbb{R}/\mathbb{Z}$. Mapping the interval to a circle (see Figure 5.1), we obtain an embedding of $\mathbb{R}/\mathbb{Z}$:

$$\varphi : \mathbb{R}/\mathbb{Z} \to \mathbb{R}, \quad \varphi(x) = (\cos(2\pi x), \sin(2\pi x)). \tag{5.6}$$

It is easily verified that the contraction of φ is given by

$$\frac{\frac{1}{2}-0}{\varphi(\pm\frac{1}{2})-\varphi(0)} = \frac{1}{4}$$

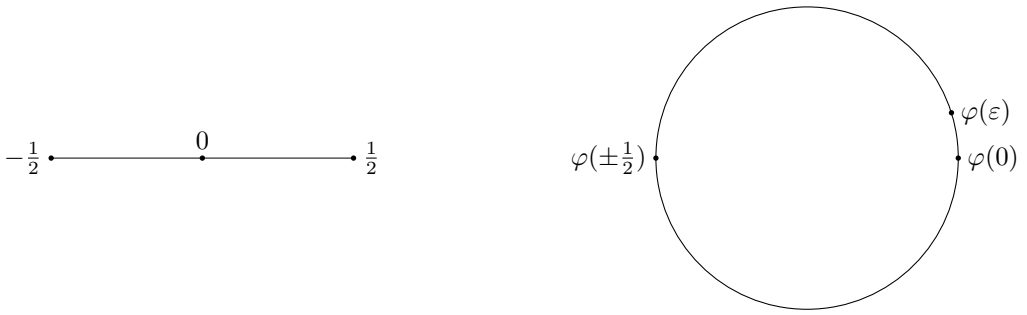


Figure 5.1: The embedding (5.6) of the torus $\mathbb{R}/\mathbb{Z}$.

and the expansion by

$$\lim_{\varepsilon \to 0} \frac{\varphi(\varepsilon) - \varphi(0)}{\varepsilon - 0} = \sqrt{2 - 2\cos(2\pi\varepsilon)} = 2\pi,$$

hence $\operatorname{dist}(\varphi) = \pi/2$. In the same way, one can show that the embedding of the standard torus $\mathbb{R}^n/\mathbb{Z}^n$ given by

$$\varphi : \mathbb{R}^n/\mathbb{Z}^n \to \mathbb{R}^{2n}, \varphi(x_1, \ldots, x_n) = (\cos 2\pi x_1, \sin 2\pi x_1, \ldots, \cos 2\pi x_n, \sin 2\pi x_n) \tag{5.7}$$

has distortion $\pi/2$.

However, in contrast to the standard torus, the metric $d_{\mathbb{R}^n/L}(x, y)$ can be highly non-Euclidean. This means that there is a family of lattices $L_n \subset \mathbb{R}^n$ such that $c_2(\mathbb{R}^n/L_n) \in \Omega(\sqrt{n})$ [KN05]. Conversely, Agarwal, Regev, Tang [ART20] showed that $c_2(\mathbb{R}^n/L) \in \mathcal{O}(\sqrt{n \log n})$ for all n-dimensional lattices L. It remains an interesting open question whether $c_2(\mathbb{R}^n/L) \in \mathcal{O}(\sqrt{n})$.

In general, studying embeddings of metric spaces has been demonstrated to be an extremely useful tool with several applications in computer science, most prominently for approximation algorithms. In the case of flat tori, a potential application might be computational lattices problems *with preprocessing* (see [FM04] for more information). For a given lattice, the preprocessing step would be to compute a Hilbert space embedding. For example, if the task is then to solve the closest vector problem for some input $x \in \mathbb{R}^n$ and a lattice $L \subset \mathbb{R}^n$, the embedding of $\mathbb{R}^n/L$ could provide helpful information, due to the

observation that

$$\min_{u \in L} |x - u| = d_{\mathbb{R}^n/L}(x, 0).$$

Euclidean embeddings of finite metric spaces (X, d) (X is a finite set, d a metric on X) have been studied from the perspective of semidefinite optimization [LLR95]. One can formulate the optimization problem of finding the least Euclidean distortion embedding of a finite metric space as semidefinite program (SDP), a linear optimization problem over the cone of positive semidefinite matrices. To see this, observe that computing the least Euclidean distortion of (X, d) amounts to solving the following optimization problem[2]

$$\begin{aligned}\min\{C \ : C \in \mathbb{R}_+, \varphi : X \to \mathbb{R}^X \text{ injective} \\ d(x,y)^2 \leq ||\varphi(x) - \varphi(y)||^2 \leq C d(x,y)^2 \text{ for all } x, y \in X\}.\end{aligned}$$

Each embedding $\varphi : X \to \mathbb{R}^X$ defines a Gram matrix $Q \in \mathbb{R}^{X \times X}$ via

$$Q_{x,y} = \varphi(x)^\mathsf{T} \varphi(y).$$

Since each positive semidefinite matrix admits a Gram matrix representation and conversely every Gram matrix is positive semidefinite, the above optimization problem can be written as an :

$$\begin{aligned}\min\{C \ : C \in \mathbb{R}_+, Q \in \mathcal{S}_+^X, \\ d(x,y)^2 \leq Q_{xx} - 2Q_{xy} + Q_{yy} \leq C d(x,y)^2 \text{ for all } x, y \in X\}.\end{aligned}$$

The SDP and its dual formulation (see Equation (7.6)) can be used to make several statements about the structure of optimal embeddings. For example, if X is a graph and d the shortest path metric, then one can show that the most expanded pairs are adjacent vertices [LM00].

In Chapter 7 we use the same strategy to compute $c_2(\mathbb{R}^n/L)$. That is, we derive an (infinite-dimensional) semidefinite program to compute the *least distortion* of $\mathbb{R}^n/L$. The semidefinite program is a generalization of the case where the underlying metric space is finite. Intuitively, this can be seen by identifying the embedding $\varphi : \mathbb{R}^n/L \to H$ with the positive semidefinite kernel

$$Q : \mathbb{R}^n/L \times \mathbb{R}^n/L \to \mathbb{C} \text{ such that } Q(x,y) = (\varphi(x), \varphi(y)) \text{ for all } x, y \in \mathbb{R}^n/L,$$

[2]The corresponding distortion is $\sqrt{C}$.

where $(\cdot,\cdot)$ denotes the inner product of H. This leads to the following formulation of $c_2(\mathbb{R}^n/L)$:

$$\begin{aligned} c_2(\mathbb{R}^n/L)^2 = \inf\{C \ : C \in \mathbb{R}_+, Q \text{ positive definite,} \\ d_{\mathbb{R}^n/L}(x,y)^2 \ &\leq \ Q(x,x) - 2\,\mathrm{Re}(Q(x,y)) + Q(y,y) \\ &\leq \ C d_{\mathbb{R}^n/L}(x,y)^2 \ \text{ for all } x,y \in \mathbb{R}^n/L\}. \end{aligned}$$

We simplify this program by using symmetry reduction techniques and derive the corresponding dual optimization problem. This symmetry reduced semidefinite program is the starting point for several interesting insights about least distortion embeddings of flat tori.

An optimal solution for the optimization problem (see (7.7)) always induces a least Euclidean distortion embedding φ of $\mathbb{R}^n/L$ into a direct product of circles $\prod_{u\in L^*} w_u S^1$, where each circle is labeled by an element of u in the dual lattice L^* and $w_u \geq 0$ is the diameter of the circle. More precisely, the embedding is given by

$$\varphi : \mathbb{R}^n/L \to \prod_{u\in L^*} w_u S^1, \quad x \mapsto \prod_{u\in L^*} w_u(\cos(2\pi u^{\mathsf{T}} x), \sin(2\pi u^{\mathsf{T}} x)).$$

This shows that the intuitively best way to embed the standard torus $\mathbb{R}^n/\mathbb{Z}^n$ fits nicely into this picture. In fact, the primal and the dual formulation for $c_2(\mathbb{R}^n/L)$ allow us to show that the embedding of $\mathbb{R}^n/\mathbb{Z}^n$ given in (5.7) is even optimal.

Another observation is that most expanded pairs only exist in the limit, i.e. these are pairs of points whose distance tends to zero. This observation also aligns with the features of embeddings of graphs with the shortest path metric: in this case, adjacent vertices, i.e. with the smallest distance are most expanded.

Furthermore, we prove that there is always a least distortion embedding of $\mathbb{R}^n/L$ that is finite-dimensional and its dimension is upper bounded by $2^{n+1}-1$. More precisely, there is always a least distortion embedding that maps $\mathbb{R}^n/L$ to a direct product of 2^n-1 circles, $\prod_{i=1}^{2^n-1} w_{u_i} S^1$.

By constructing a dually feasible solution, we give a simple proof for a constant factor improvement of the previously best known lower

bound [KN05, HR13] that

$$c_2(\mathbb{R}^n/L) \geq \frac{\pi\lambda(L^*)\mu(L)}{\sqrt{n}},$$

where L^* is the dual lattice of L, the parameter $\lambda(L^*)$ is the length of the shortest vector in L^* and $\mu(L)$ is the circumradius of L (see Chapter 7 for precise definitions).

Finally, we construct optimal embeddings for all tori $\mathbb{R}^n/L$ where L is a 2-dimensional lattice. We achieve this by simply finding a feasible primal and feasible dual solution that yield the same value.

In summary, we provide a new approach to compute (optimal) distortion embeddings of flat tori. However, it remains an interesting open question whether this approach can be exploited from an algorithmic point of view to find new algorithms for computational lattice problems.

Chapter 6

Critical even unimodular lattices in the Gaussian core model

Critical even unimodular lattices in the Gaussian core model

About this section

The following text has been previously published as:

> Arne Heimendahl, Aurelio Marafioti, Antonia Thiemeyer, Frank Vallentin, Marc Christian Zimmermann. "Critical Even Unimodular Lattices in the Gaussian Core Model." In: *International Mathematics Research Notices*, 2022, rnac164, `https://doi.org/10.1093/imrn/rnac164`

Changes from the journal version are limited to typesetting and notation. These changes were performed to match the rest of this dissertation.

All authors contributed equally to this work. AH was particularly involved in working out the results of Section 6.4.

Abstract

We consider even unimodular lattices which are critical for potential energy with respect to Gaussian potential functions in the manifold of lattices having point density 1. All even unimodular lattices up to dimension 24 are critical. We show how to determine the Morse index in these cases. While all these lattices are either local minima or

saddle points, we find lattices in dimension 32 which are local maxima. Also starting from dimension 32 there are non-critical even unimodular lattices.

6.1 Introduction

Let $L \subseteq \mathbb{R}^n$ be an n-dimensional lattice (a discrete subgroup of $\mathbb{R}^n$ of full rank). Let $f : (0, \infty) \to \mathbb{R}$ be a nonnegative function, then the f-potential energy of L is defined as

$$\mathcal{E}(f, L) = \sum_{x \in L \setminus \{0\}} f(\|x\|^2).$$

In this paper we are mainly interested in Gaussian potential functions $f_\alpha(r) = e^{-\alpha r}$ with $\alpha > 0$. Point configurations which interact via such a Gaussian potential function are referred to as the Gaussian core model. They are natural physical systems (see [Sti]) and they are mathematically quite general. By Bernstein's theorem (see [Wid41, Theorem 12b, page 161]), Gaussian potential functions span the convex cone of completely monotonic functions (C^∞-functions f with $(-1)^k f^{(k)} \geq 0$ for all $k \in \mathbb{N}$) of squared Euclidean distance.

We are interested in a local analysis of the function $L \mapsto \mathcal{E}(f_\alpha, L)$ when L varies in the manifold of rank n lattices having point density 1, which means that the number of lattice points per unit volume equals 1. In particular, we want to understand which even unimodular lattices are critical points in the Gaussian core model and which type they have.

Recall that a lattice L is called unimodular if it coincides with its dual lattice, which is defined as

$$L^* = \{y \in \mathbb{R}^n : x \cdot y \in \mathbb{Z} \text{ for all } x \in L\},$$

where $x \cdot y$ denotes the standard inner product of $x, y \in \mathbb{R}^n$. The lattice L is called even if for every lattice vector $x \in L$ the inner product $x \cdot x$ is an even integer. It is well-known that in a given dimension the number of even unimodular lattices is finite and that they exist only in dimensions which are divisible by 8. Furthermore, dimensions 8

and 24 seem to be very special. Cohn, Kumar, Miller, Radchenko and Viazovska [CKM+22] proved that the E_8 root lattice in dimension 8 and the Leech lattice Λ_{24} in dimension 24 are universally optimal point configurations in their dimensions. This means that they minimize f-potential energy for all point configurations having density 1 in their dimensions (not only for lattices) and for all completely monotonic functions of squared Euclidean distance.

6.1.1 Structure of the paper and main results

In Section 6.5 we present our concrete results. Here we summarize the phenomena which occur.

Dimension 8

Section 6.5.1: In dimension 8 the E_8 root lattice is the only even unimodular lattice in dimension 8 as observed by Mordell [Mor38]. It is universally optimal. In particular, it is a local minimum for f_α-potential energy. This was first proved by Sarnak and Strömbergsson [SS06], see also Coulangeon [Cou06].

Dimension 16

Section 6.5.2: In dimension 16 there are two even unimodular lattices D_{16}^+ and $E_8 \perp E_8$, first classified by Witt [Wit41]. Both of them are critical and we show that D_{16}^+ is a local minimum for f_α-potential energy whenever α is large enough and that $E_8 \perp E_8$ is a saddle point whenever α is large enough. Our numerical computations strongly suggest that $E_8 \perp E_8$ is a saddle point for all values of α.

Dimension 24

Section 6.5.3: Apart from the universally optimal Leech lattice there are 23 further even unimodular lattices in dimension 24. They were first classified by Niemeier [Nie73]. Again they are all critical. We show how to determine their Morse index. We always find either local minima or saddle points.

Dimension 32

Section 6.5.4: It is known that there are more than 80 millions even unimodular lattices in dimension 32; cf. Serre [Ser73]. A complete classification has not been achieved yet. We show that not all of them are critical. We also show that there exist local *maxima* for f_α-potential energy. This existence of local maxima answers a question of Regev and Stephens-Davidowitz [RSD17] which arose in their proof strategy of the reverse Minkowski theorem; see also the exposition [Bos18] by Bost for a broad perspective. A similar phenomenon, a local maximum for the covering density of a lattice, was earlier found by Dutour Sikirić, Schürmann, and Vallentin [SSV12].

Proof techniques

To prove these results we make use of the theory of lattices and codes, especially spherical designs, theta series with spherical coefficients, and root systems. We recall these tools in Section 6.2. In Section 6.3 we describe our strategy which is based on the explicit computation of the signature of the Hessian of the function $L \mapsto \mathcal{E}(f_\alpha, L)$. To work out this strategy it is necessary to explicitly compute the eigenvalues of a symmetric matrix which is parametrized by root systems. This is done in Section 6.4.

6.2 Toolbox

In this section we introduce the tools we shall apply later in this paper. For more information we refer to the standard literature on lattices and codes, in particular to Conway and Sloane [CS88], Ebeling [Ebe94], Serre [Ser73], Venkov [Ven01], Nebe [Neb13]. Readers familiar with lattices and codes might like to skip immediately to the next section.

6.2.1 Spherical designs

A finite set X on the sphere of radius r in $\mathbb{R}^n$ denoted by $S^{n-1}(r)$ is called a spherical t-design if

$$\int_{S^{n-1}(r)} p(x)\,dx = \frac{1}{|X|}\sum_{x\in X} p(x)$$

holds for every polynomial p of degree up to t. Here we integrate with respect to the rotationally invariant probability measure on $S^{n-1}(r)$.

If X forms a spherical 2-design, then

$$\sum_{x\in X} xx^{\mathsf{T}} = \frac{r^2|X|}{n} I_n, \tag{6.1}$$

holds, where I_n denotes the identity matrix with n rows/columns.

A polynomial $p \in \mathbb{R}[x_1, \ldots, x_n]$ is called harmonic if it vanishes under the Laplace operator

$$\Delta p = \sum_{i=1}^{n} \frac{\partial^2 p}{\partial x_i^2} = 0.$$

We denote the space of homogeneous harmonic polynomials of degree k by Harm_k. One can uniquely decompose every homogeneous polynomial p of even degree k

$$p(x) = p_k(x) + \|x\|^2 p_{k-2}(x) + \|x\|^4 p_{k-4}(x) + \cdots + \|x\|^k p_0(x) \tag{6.2}$$

with $p_d \in \mathrm{Harm}_d$ and $d = 0, 2, \ldots, k$.

We can characterize that X is a spherical t-design by saying that the sum $\sum_{x\in X} p(x)$ vanishes for all homogeneous harmonic polynomials p of degree $1, \ldots, t$.

In the following we shall need the following technical lemma.

Lemma 6.2.1. *Let H be a symmetric matrix with trace zero. The homogeneous polynomial*

$$p_H(x) = (x^{\mathsf{T}} H x)^2 = H[x]^2$$

of degree four decomposes as in (6.2)

$$p_H(x) = p_{H,4}(x) + \|x\|^2 p_{H,2}(x) + \|x\|^4 p_{H,0}(x)$$

with $p_{H,d} \in \mathrm{Harm}_d$ *and*

$$p_{H,4}(x) = p_H(x) - \|x\|^2 \frac{4}{4+n} H^2[x] + \|x\|^4 \frac{2}{(4+n)(2+n)} \operatorname{Tr} H^2$$

and

$$p_{H,0}(x) = \frac{2}{(2+n)n} \operatorname{Tr} H^2.$$

Proof. As a consequence of Euler's formula we have for a general harmonic polynomial $q \in \mathrm{Harm}_d$

$$\Delta \|x\|^2 q = (4d + 2n)q + \|x\|^2 \Delta q = (4d + 2n)q,$$

and inductively

$$\Delta \|x\|^{2(k+1)} q = (k+1)(4k + 4d + 2n)\|x\|^{2k} q, \tag{6.3}$$

see for example [Sim15, Lemma 3.5.3][1].

Using (6.2) we get

$$\begin{aligned} \Delta p_H &= \Delta p_{H,4} + (8 + 2n)p_{H,2} + \|x\|^2 \Delta p_{H,2} + \Delta \|x\|^4 p_{H,0} \\ &= (8 + 2n)p_{H,2} + 2(4 + 2n)\|x\|^2 p_{H,0}. \end{aligned}$$

Applying the Laplace operator another time yields

$$\Delta^2 p_H = 8n(n+2)p_{H,0}.$$

On the other hand, one can compute $\Delta^2 p_H$ directly. We have

$$H[x] = \sum_{i=1}^{n} \sum_{j=1}^{n} H_{ij} x_i x_j$$

and therefore

$$\Delta H[x] = 2 \sum_{i=1}^{n} H_{ii} = 2 \operatorname{Tr} H.$$

Using the product formula for the Laplace operator and the symmetry of H we get

$$\Delta p_H = \Delta H[x]^2 = 2(H[x]\Delta H[x] + \nabla H[x] \cdot \nabla H[x]) = 4(\operatorname{Tr} H)H[x] + 8H^2[x].$$

[1] The factor 2 in (3.5.11) is wrong in [Sim15]; it should be 1.

Therefore

$$\Delta^2 p_H = 8(\operatorname{Tr} H)^2 + 16 \operatorname{Tr} H^2$$

and so

$$p_{H,0} = \frac{2}{n(n+2)} \operatorname{Tr} H^2,$$

where the last equation follows from $\operatorname{Tr} H = 0$.

We already computed

$$\Delta \|x\|^4 p_{H,0} = 2(4+2n)\|x\|^2 p_{H,0} = \frac{8}{n} \operatorname{Tr} H^2 \|x\|^2.$$

Now we determine $p_{H,2}$ when $\operatorname{Tr} H = 0$:

$$(8+2n)p_{H,2} = \Delta p_H - \|x\|^2 \frac{8}{n} \operatorname{Tr} H^2 = 8H^2[x] - \|x\|^2 \frac{8}{n} \operatorname{Tr} H^2.$$

Finally we get $p_{H,4}$:

$$p_{H,4} = p_H - \|x\|^2 \frac{4}{4+n} H[x]^2 + \|x\|^4 \frac{2}{(4+n)(2+n)} \operatorname{Tr} H^2. \qquad \square$$

6.2.2 Theta series with spherical coefficients

We will make use of theta series with spherical coefficients. Let $L \subseteq \mathbb{R}^n$ be an even unimodular lattice and let p be a harmonic polynomial (sometimes also called spherical polynomial).

We define the theta series of L with spherical coefficients given by p by

$$\Theta_{L,p}(\tau) = \sum_{x \in L} p(x) e^{\pi i \tau \|x\|^2} = \sum_{x \in L} p(x) q^{\frac{1}{2}\|x\|^2},$$

where τ lies in the upper half plane $\{z \in \mathbb{C} : \operatorname{Im}(z) > 0\}$ and where $q = e^{2\pi i \tau}$.

If $p = 1$ we also write Θ_L instead of $\Theta_{L,p}$. For $r \geq 0$ we define

$$L(r^2) = \{x \in L : x \cdot x = r^2\}.$$

The set $L(r^2)$ is called a shell of L if it is not empty. Then

$$\Theta_L(\tau) = \sum_{m=0}^{\infty} a_m q^m \quad \text{with} \quad a_m = |L(2m)|.$$

The theta series of L is related to its f_α-potential energy through

$$\mathcal{E}(f_\alpha, L) = \Theta_L(\alpha i/\pi) - 1.$$

Using the Poisson summation formula one sees that

$$\Theta_L(iy) = y^{-n/2}\Theta_{L^*}(i/y) \text{ for } y > 0.$$

In particular, when $L = L^*$ it is sufficient to consider Gaussian potentials with $\alpha \geq \pi$.

If p is a homogeneous harmonic polynomial of degree k, then $\Theta_{L,p}$ is a modular form (for the full modular group $\mathrm{SL}_2(\mathbb{Z})$) of weight $n/2+k$. When $k > 1$ then $\Theta_{L,p}$ is a cusp form. We only need that modular forms form a graded ring which is isomorphic to the polynomial ring $\mathbb{C}[E_4, E_6]$ in the (normalized) Eisenstein series

$$E_4(\tau) = 1 + 240q + 2160q^2 + 6720q^3 + \cdots,$$

and

$$E_6(\tau) = 1 - 504q - 16632q^2 - 122976q^3 - \cdots,$$

where the weight of the monomial $E_4^\alpha E_6^\beta$ is $4\alpha + 6\beta$. Generally, the normalized Eisenstein series are given by

$$E_k(\tau) = 1 - \frac{2k}{B_k}\sum_{m=1}^{\infty}\sigma_{k-1}(m)q^m \quad \text{for } k \geq 4,$$

where B_k is the k-th Bernoulli number and where $\sigma_{k-1}(m) = \sum_{d|m} d^{k-1}$ is the sum of the $(k-1)$-th powers of positive divisors of m. The space of cusp forms is a principal ideal of the polynomial ring $\mathbb{C}[E_4, E_6]$ generated by the modular discriminant

$$\Delta = \frac{1}{1728}(E_4^3 - E_6^2) = 0 + q - 24q^2 + 252q^3 \pm \cdots,$$

which has weight 12.

It is a standard fact that the cardinality $a_m = |L(2m)|$ of the shells is asymptotically bounded, when m tends to infinity, by

$$a_m = -\frac{n}{B_{n/2}}\sigma_{n/2-1}(m) + O(m^{n/4}),$$

but in this paper we shall need a bound with explicit constants.

For this we we will use the following explicit bound by Jenkins and Rouse [JR11] which relies on Deligne's proof of the Weil conjectures: Let $f(\tau) = \sum_{m=1}^{\infty} a_m q^m$ be a cusp form of weight k, let ℓ be the dimension of the space of cusp forms of weight k, then

$$|a_m| \leq \sqrt{\log(k)} \left(11 \cdot \sqrt{\sum_{r=1}^{\ell} \frac{|a_r|^2}{r^{k-1}}} + \frac{e^{18.72}(41.41)^{k/2}}{k^{(k-1)/2}} \cdot \left| \sum_{r=1}^{\ell} a_r e^{-7.288r} \right| \right) \cdot d(m) m^{\frac{k-1}{2}}, \tag{6.4}$$

where $d(m)$ is the number of divisors of m.

The following simple estimate will be helpful several times.

Lemma 6.2.2. *For $j \geq k/(2\alpha)$ we have*

$$\sum_{m=j}^{\infty} m^k e^{-2\alpha m} \leq j^k e^{-2\alpha j} + (2\alpha)^{-(k+1)} \Gamma(k+1, 2\alpha j), \tag{6.5}$$

where

$$\Gamma(s,x) = \int_x^{\infty} t^{s-1} e^{-t} dt$$

is the incomplete gamma function.

As for fixed s and large x

$$\Gamma(s,x) \sim x^{s-1} e^{-x} \left(1 + \frac{s-1}{x} + \frac{(s-1)(s-2)}{x^2} + \cdots \right)$$

we see that (6.5) tends to zero for large α and fixed j and k.

Proof. The function $m \mapsto m^k e^{-2\alpha m}$ is monotonically decreasing for $m \geq k/(2\alpha)$. So we can apply the integral test

$$\sum_{m=j}^{\infty} m^k e^{-2\alpha m} \leq j^k e^{-2\alpha j} + \int_j^{\infty} m^k e^{-2\alpha m} \, dm.$$

Now using the definition of the incomplete gamma function after a change of variables yields the lemma. □

6.2.3 Root systems

The shell $L(2)$ is called the root system of the even unimodular lattice L, its elements are called roots. Witt classified in 1941 the possible root systems: These are orthogonal direct sums of the irreducible root systems A_n ($n \geq 1$), D_n ($n \geq 4$), E_6, E_7 and E_8. The rank of a root system is the dimension of the vector space it spans. Let $e_1, \ldots, e_{n+1}$ be the standard basis for $\mathbb{R}^{n+1}$. The root system A_n is defined as

$$\{\pm(e_i - e_j) : 1 \leq i < j \leq n+1\}.$$

The root system A_n has rank n, but lies in $\mathbb{R}^{n+1}$. It spans the vector space $\mathbb{R}^{n+1} \cap \mathbb{R}(1, \ldots, 1)^\perp \cong \mathbb{R}^n$. In the following we will consider A_n as a subset in $\mathbb{R}^n$. The root system D_n is defined as

$$D_n = \{\pm(e_i \pm e_j) : 1 \leq i < j \leq n\}.$$

Furthermore

$$E_8 = D_8 \cup \left\{ \frac{e_1 \pm \cdots \pm e_8}{2} \right\},$$

where we restrict the last set to all sums having an even number of minus signs, and

$$E_7 = E_8 \cap \mathbb{R}(e_7 - e_8)^\perp \quad \text{and} \quad E_6 = E_7 \cap \mathbb{R}(e_6 - e_7)^\perp.$$

All irreducible root systems form spherical 2-designs, and we have even spherical 4-designs for A_1, A_2, D_4, E_6, E_7, and E_8.

Let R be a root system. Let $\sigma(x) = I_n - xx^\mathsf{T}$ be the reflection at the hyperplane perpendicular to x. For all $x, y \in R$ we have $\sigma(x)y \in R$, so that R is invariant under the reflection $\sigma(x)$. The group $W(R)$ generated by all reflections $\sigma(x)$, with $x \in R$, is called Weyl group of the root system.

The Coxeter number h of a root system R with rank n is defined as $|R|/n$, the number of roots per dimension. For a root $r \in R$ we denote by n_0 the number of roots $r' \in R$ with $r \cdot r' = 0$ and by n_1 the number of roots $r' \in R$ with $r \cdot r' = 1$. These numbers n_0, n_1 do not depend on r when R is irreducible.

We summarize some properties of the irreducible root systems in Table 6.1.

name	rank	$\lvert R\rvert$	n_0	n_1	h	$\lvert W\rvert$
A_n	$n \geq 1$	$n(n+1)$	$(n-1)(n-2)$	$2(n-1)$	$n+1$	$(n+1)!$
D_n	$n \geq 4$	$2n(n-1)$	$2(n^2-5n+7)$	$4(n-2)$	$2(n-1)$	$2^{n-1}n!$
E_6	6	72	30	20	12	$2^7 3^4 5$
E_7	7	126	60	32	18	$2^{10} 3^4 57$
E_8	8	240	126	56	30	$2^{14} 3^5 5^2 7$

Table 6.1: Some properties of the irreducible root systems.

6.3 Strategy

We compute the gradient and Hessian of $L \mapsto \mathcal{E}(f_\alpha, L)$ at even unimodular lattices. For this it is convenient to parametrize the manifold of rank n lattices having point density 1 by positive definite quadratic forms of determinant 1.

The gradient and the Hessian of $\mathcal{E}(f_\alpha, L)$ at L were computed by Coulangeon and Schürmann [CS12, Lemma 3.2]. Let H be a symmetric matrix having trace zero (lying in the tangent space of the identity matrix). We use the notation $H[x] = x^\mathsf{T} H x$ and we equip the space of symmetric matrices $\mathcal{S}^n$ with the inner product $\langle A, B\rangle = \operatorname{Tr}(AB)$, where $A, B \in \mathcal{S}^n$. The gradient is given by

$$\langle \nabla \mathcal{E}(f_\alpha, L), H\rangle = -\alpha \sum_{x \in L\setminus\{0\}} H[x] e^{-\alpha\|x\|^2}. \tag{6.6}$$

Now a sufficient condition for L being a critical point is that all shells of L form spherical 2-designs. Indeed, we group the sum in (6.6) according to shells, giving

$$\langle \nabla \mathcal{E}(f_\alpha, L), H\rangle = -\alpha \sum_{r>0} e^{-\alpha r^2} \sum_{x \in L(r^2)} H[x].$$

Then for $r > 0$ every summand

$$\sum_{x \in L(r^2)} H[x] = \left\langle H, \sum_{x \in L(r^2)} xx^\mathsf{T} \right\rangle = \frac{r^2 |X|}{n} \operatorname{Tr}(H) = 0$$

vanishes because of (6.1) and because H is traceless. Hence, L is critical.

This sufficient condition is fulfilled for all even unimodular lattices in dimensions 8, 16, and 24. This fact can be deduced from the theory of theta functions with spherical coefficients and modular forms as first observed by Venkov [Ven80]. In dimension 32 this is no longer fulfilled in general but we can identify cases where it is.

The Hessian is the quadratic form

$$\nabla^2 \mathcal{E}(f_\alpha, L)[H] = \alpha \sum_{x \in L \setminus \{0\}} e^{-\alpha \|x\|^2} \left(\frac{\alpha}{2} H[x]^2 - \frac{1}{2} H^2[x] \right). \tag{6.7}$$

Again grouping the sum according to shells we get

$$\nabla^2 \mathcal{E}(f_\alpha, L)[H] = \alpha \sum_{r>0} e^{-\alpha r^2} \sum_{x \in L(r^2)} \left(\frac{\alpha}{2} H[x]^2 - \frac{1}{2} H^2[x] \right). \tag{6.8}$$

So it remains to determine the two sums

$$\sum_{x \in L(r^2)} H[x]^2 \quad \text{and} \quad \sum_{x \in L(r^2)} H^2[x]. \tag{6.9}$$

The second sum is easy to compute when $L(r^2)$ forms a spherical 2-design. In this case we have by (6.1)

$$\sum_{x \in L(r^2)} H^2[x] = \left\langle H^2, \sum_{x \in L(r^2)} x x^\mathsf{T} \right\rangle = \langle H^2, \frac{r^2 |L(r^2)|}{n} I_n \rangle = \frac{r^2 |L(r^2)|}{n} \operatorname{Tr} H^2. \tag{6.10}$$

The first sum is only easy to compute when $L(r^2)$ forms even a spherical 4-design. Then (see [Cou06, Proposition 2.2] for the computation)

$$\sum_{x \in L(r^2)} H[x]^2 = \frac{r^4 |L(r^2)|}{n(n+2)} 2 \operatorname{Tr} H^2. \tag{6.11}$$

Together, when all shells form spherical 4-designs, the Hessian (6.7) simplifies to

$$\nabla^2 \mathcal{E}(f_\alpha, L)[H] = \frac{\operatorname{Tr} H^2}{n(n+2)} \sum_{r>0} |L(r^2)| \alpha r^2 \left(\alpha r^2 - (n/2+1) \right) e^{-\alpha r^2}. \tag{6.12}$$

Therefore, every H with Frobenius norm $\langle H, H \rangle = \operatorname{Tr} H^2 = 1$ is mapped to the same value, which implies that all the eigenvalues of the Hessian coincide.

Sarnak and Strömbergsson [SS06], see also Coulangeon [Cou06], showed that for $L = E_8, \Lambda_{24}$ the Hessian $\nabla^2\mathcal{E}(f_\alpha, L)[H]$ is positive for all $\alpha > 0$ which implies that E_8, Λ_{24} are local minima among lattices, for all completely monotonic potential functions of squared Euclidean distance[2].

The case when all shells form spherical 2-designs but not spherical 4-designs requires substantially more work. This is our main technical contribution. Then the Hessian has more than only one eigenvalue. We determine these eigenvalues up to dimension 32 by considering the root system of L, that is the shell $L(2)$. Here the quadratic form

$$Q[H] = \sum_{x \in L(2)} H[x]^2 \tag{6.13}$$

will play a crucial role.

Indeed, consider again the first sum $\sum_{x\in L(r^2)} H[x]^2$ in (6.9). We decompose the polynomial $p_H(x) = H[x]^2$ into its harmonic components as in Lemma 6.2.1 and get

$$\sum_{x\in L(r^2)} p_H(x) = \sum_{x\in L(r^2)} p_{H,4}(x) + r^2 \sum_{x\in L(r^2)} p_{H,2}(x) + r^4 \sum_{x\in L(r^2)} p_{H,0}(x).$$

Here the first sum equals

$$\sum_{x\in L(r^2)} p_{H,4}(x) = \sum_{x\in L(r^2)} H[x]^2 - r^4 \frac{2}{(2+n)n} |L(r^2)| \operatorname{Tr} H^2,$$

where we used Lemma 6.2.1 and (6.10). The second sum vanishes because $L(r^2)$ is a spherical 2-design and the third summand equals

$$r^4 \sum_{x\in L(r^2)} p_{H,0}(x) = r^4 \frac{2}{(2+n)n} |L(r^2)| \operatorname{Tr} H^2.$$

We make use of theta series with spherical coefficients to determine the first sum $\sum_{x\in L(r^2)} p_{H,4}(x)$ explicitly: $\Theta_{L,p_{H,4}}$ is a cusp form of weight $n/2+4$. In dimension 16, 24, and 32 there is (up to scalar multiplication) only one cusp form of weight $n/2+4$. This is, respectively, Δ,

[2]This was one motivation for Cohn, Kumar, Miller, Radchenko, and Viazovska [CKM+22] to prove their far stronger, global result.

$E_4\Delta$ and $E_4^2\Delta$. Their q-expansions $\sum_{m=0}^{\infty} b_m q^m$ all start by $0 + 1 \cdot q$. Therefore, by equating coefficients,

$$\Theta_{L,p_{H,4}}(\tau) = \sum_{r>0} \sum_{x \in L(r^2)} p_{H,4}(x) q^{\frac{1}{2}r^2} = c \sum_{m=0}^{\infty} b_m q^m$$

with

$$c = \sum_{x \in L(2)} H[x]^2 - \frac{8}{(2+n)n} |L(2)| \operatorname{Tr} H^2.$$

For $r^2 = 2m$ it follows

$$\sum_{x \in L(r^2)} H[x]^2 = c b_m + 4m^2 \frac{2}{(2+n)n} |L(2m)| \operatorname{Tr} H^2.$$

Hence, we only need to compute the eigenvalues of (6.13) to determine the signature of the Hessian. When talking about eigenvalues of Q, we refer to the eigenvalues of the Gram matrix with entries $b_Q(G_i, G_j)$, where $b_Q : \mathcal{S}^n \times \mathcal{S}^n \to \mathbb{R}$ is the induced bilinear form

$$b_Q(G, H) = \sum_{x \in L(2)} G[x] H[x] \tag{6.14}$$

and (G_i) is an orthonormal basis of the space $\mathcal{S}^n$ with respect to the inner product $\langle \cdot, \cdot \rangle$. If H is an eigenvector with eigenvalue λ, we have

$$\sum_{x \in L(2)} H[x]^2 = \lambda \operatorname{Tr} H^2.$$

Now let us put everything together.

Theorem 6.3.1. *Let L be an even unimodular lattice in dimension $n \leq 32$. Let*

$$\Theta_L(\tau) = \sum_{m=0}^{\infty} a_m q^m \quad \text{with } a_m = |L(2m)|$$

be the theta series of L and let $\sum_{m=1}^{\infty} b_m q^m$ be the cusp form of weight $n/2+4$ with $b_1 = 1$. Then all the eigenvalues of the Hessian $\nabla^2 \mathcal{E}(f_\alpha, L)$ are given by

$$\begin{aligned} &\frac{1}{n(n+2)} \sum_{m=1}^{\infty} \left(b_m \frac{\alpha^2}{2} (\lambda n(n+2) - 8a_1) \right) e^{-2\alpha m} \\ &+ \frac{1}{n(n+2)} \sum_{m=1}^{\infty} \left(a_m 2\alpha m \left(2\alpha m - (n/2+1) \right) \right) e^{-2\alpha m}, \end{aligned} \tag{6.15}$$

where λ is an eigenvalue of (6.13).

Note that this theorem also includes the case when all shells of L form spherical 4-designs like in (6.12) because of (6.11). In this case and when the parameter α is large enough, then (6.12) is strictly positive, which shows that L is a local minimum for f_α-potential energy.

Similarly, because the growth of a_m and b_m is polynomial in m and because of the estimate provided in Lemma 6.2.2, we see that the first summand, $m = 1$,

$$\frac{1}{n(n+2)}\left(\frac{\alpha^2}{2}(\lambda n(n+2)) - 2a_1\alpha(n/2+1)\right)e^{-2\alpha}$$

dominates (6.15) for large α. In particular, for large α, the first summand is strictly positive if λ is strictly positive and the first summand is strictly negative if λ vanishes and if $a_1 \neq 0$. As the quadratic form (6.13) is a non-trivial sum of squares, the eigenvalues cannot be strictly negative and some eigenvalue is always strictly positive. From this consideration we get:

Corollary 6.3.2. *Let L be an even unimodular lattice in dimension $n \leq 32$ which is critical for f_α-potential energy. For all large enough α the lattice L is a local minimum if and only if all eigenvalues of* (6.13) *are strictly positive. If one eigenvalue of* (6.13) *vanishes and if $|L(2)| > 0$, then L is a saddle point for all large enough α.*

6.4 Eigenvalues of (6.13)

In this section we shall compute the eigenvalues of the quadratic form (6.13) $Q[H] = \sum_{x\in R} H[x]^2$, where we write $R = L(2)$ for the root system of the lattice.

6.4.1 Irreducible root systems

First we consider the case when R is an irreducible root system of type A, D, or E.

Theorem 6.4.1. *Let R be an irreducible root system of type A, D, or E. The quadratic form $Q[H] = \sum_{x \in R} H[x]^2$ has the following eigenvalues:*

root system	eigenvalue	multiplicity
$A_n,\ n \geq 1$	$4h = 4(n+1)$	1
	$2(n+1)$	n, *for* $n \geq 2$
	4	$n(n-1)/2 - 1$, *for* $n \geq 2$
$D_n,\ n \geq 4$	$4h = 8(n-1)$	1
	$4(n-2)$	$n-1$
	8	$n(n-1)/2$
E_6	$4h = 48$	1
	12	20
E_7	$4h = 72$	1
	16	27
E_8	$4h = 120$	1
	24	35

We will embed the proof of Theorem 6.4.1 in the framework of representation theory.[3] The Weyl group W of the root system R acts on the space of symmetric matrices $\mathcal{S}^n$ by conjugation

$$\begin{aligned} W \times \mathcal{S}^n &\to \mathcal{S}^n \\ (S, H) &\mapsto SHS^\mathsf{T}. \end{aligned}$$

This turns $(\mathcal{S}^n, \langle \cdot, \cdot \rangle)$ into a unitary representation of W, meaning that the action of W preserves the inner product $\langle \cdot, \cdot \rangle$.

Then the bilinear form b_Q, defined in (6.14), is invariant under the action of the Weyl group W, that is $b_Q(SGS^\mathsf{T}, SHS^\mathsf{T}) = b_Q(G, H)$ for all $S \in W$. Due to the Riesz representation theorem, there is a linear map $T : \mathcal{S}^n \to \mathcal{S}^n$ such that

$$b_Q(G, H) = \langle G, T(H) \rangle$$

and the eigenvalues of the Gram matrix of b_Q coincide with the eigenvalues of T. Since b_Q is invariant under the action of W, the map T

[3] In the following we apply concepts of unitary representations over the complex numbers, but note that all representations involved can in fact be defined over the reals.

commutes with the action of W, i.e.

$$T(SHS^{\mathsf{T}}) = ST(H)S^{\mathsf{T}} \quad \text{for all } S \in W, \tag{6.16}$$

hence, T is intertwining the representation $(\mathcal{S}^n, \langle\cdot,\cdot\rangle)$ of the Weyl group W with itself.

Instead of only considering the specific map T above, we determine the common eigenspaces of all intertwiners that intertwine the representation on $\mathcal{S}^n$ with itself. As these eigenspaces will turn out to be inequivalent, Schur's lemma implies that these eigenspaces are exactly the pairwise orthogonal, irreducible, W-invariant subspaces of $\mathcal{S}^n$.

6.4.2 Peter-Weyl theorem for irreducible root systems

This gives rise to Theorem 6.4.2, which is a *Peter-Weyl theorem* for the representation $(\mathcal{S}^n, \langle\cdot,\cdot\rangle)$ of the Weyl group W of an irreducible root system.

To state the theorem, we need to fix some notation, based on the definition of root systems in Section 6.2.3. We consider A_n as a root system in $\mathbb{R}^n$ and, by slight abuse of notation, we write $e_i - e_j$ for the corresponding root in $\mathbb{R}^n$. Moreover, define the symmetric bilinear operator $M : \mathbb{R}^n \times \mathbb{R}^n \to \mathcal{S}^n$ by

$$M(x,y) = xy^{\mathsf{T}} + yx^{\mathsf{T}}.$$

The action of the Weyl group on M is given by

$$SM(x,y)S^{\mathsf{T}} = M(Sx, Sy), \qquad S \in W.$$

Furthermore, set

$$M(x) = \frac{1}{2}M(x,x) = xx^{\mathsf{T}}$$

and

$$P_i = \sum_{j\in\{1,\dots,n+1\}\setminus\{i\}} M(e_i - e_j) - 2I_n.$$

Theorem 6.4.2 (Peter-Weyl for irreducible root systems)**.** *The space of symmetric matrices can be decomposed into the following W-invariant, irreducible, inequivalent subspaces:*

(i) For $R = A_n,\ n \geq 2$

$$\mathcal{S}^n = \operatorname{span}\{I_n\} \perp U_1(A_n) \perp U_2(A_n),$$

where

$$\begin{aligned} U_1(A_n) &= \operatorname{span}\{M(x,y) : x, y \in A_n,\ x \cdot y = 0\}, \\ U_2(A_n) &= \operatorname{span}\{P_i : i = 1, \ldots, n+1\}. \end{aligned}$$

(ii) For $R = D_n,\ n \geq 5$

$$\mathcal{S}^n = \operatorname{span}\{I_n\} \perp U_1(D_n) \perp U_2(D_n),$$

where

$$U_1(D_n) = \{M \in \mathcal{S}^n : M_{ii} = 0,\ 1 \leq i \leq n\}. \tag{6.17}$$

and

$$U_2(D_n) = \{\operatorname{diag}(d_1, \ldots, d_n) : d_1, \ldots, d_n \in \mathbb{R},\ d_1 + \cdots + d_n = 0\}. \tag{6.18}$$

For $n = 4$ *the space* $U_1(D_4)$ *further splits into two irreducible subspaces*

$$\begin{aligned} U_1(D_4) \ = \ & \left\{ \begin{pmatrix} 0 & a & b & c \\ a & 0 & c & b \\ b & c & 0 & a \\ c & b & a & 0 \end{pmatrix} \ : \ a, b, c \in \mathbb{R} \right\} \\ \perp \ & \left\{ \begin{pmatrix} 0 & a & b & -c \\ a & 0 & c & -b \\ b & c & 0 & -a \\ -c & -b & -a & 0 \end{pmatrix} \ : \ a, b, c \in \mathbb{R} \right\}. \end{aligned} \tag{6.19}$$

(iii) For $R \in \{E_6, E_7, E_8\}$

$$\mathcal{S}^n = \operatorname{span}\{I_n\} \perp \mathcal{T}_0^n,$$

where $\mathcal{T}_0^n$ *is the space of traceless symmetric* $n \times n$ *matrices.*

Remark 6.4.3. *The proofs of (i) and (ii) will be based on the representation theory of the symmetric group*[4] *(see [FH91, Chapter 4] for details). In fact, the decompositions are immediate consequences of the representation theory of the symmetric group, most of the work lies in the explicit description of the irreducible subrepresentations, as we need these, for the explicit calculation of the eigenvalues in Theorem 6.4.1.*

We will give an elementary proof of (iii) in Section 6.4.5. However, as one of the anonymous referees pointed out, this could also be done by computing the explicit characters of the representation, as it was already done in the literature. See [Fra51] for the case E_6 and E_7, and [Fra70], for the case E_8.

The main ingredient is a decomposition formula for a representation of $\mathfrak{S}_{n+1}$, the symmetric group on $n+1$ symbols. We write

$$U = \operatorname{span}\{e\},$$

where e is the all ones vector, for the trivial representation and

$$V_{n+1} = \left\{ v \in \mathbb{R}^{n+1} \ : \ \sum_{i=1}^{n+1} v_i = 0 \right\} = U^{\perp} \tag{6.20}$$

for the standard representation of $\mathfrak{S}_{n+1}$. Clearly U and V_{n+1} are orthogonal as representations. Furthermore, both are irreducible: they are the cases of a standard principle to construct the irreducible representations of $\mathfrak{S}_{n+1}$ via Young symmetrizers, which give a one-to-one correspondence between partitions of $n+1$ and irreducible representations of $\mathfrak{S}_{n+1}$ [FH91, Theorem 4.3].

One then obtains the decomposition[5]

$$\operatorname{Sym}^2(V_{n+1}) \cong U \oplus V_{n+1} \oplus V_{((n+1)-2,2)}, \tag{6.21}$$

where $V_{((n+1)-2,2)}$ is another irreducible representation[6] of $\mathfrak{S}_{n+1}$.

[4] The authors would like to thank one of the anonymous referees for the suggestion and a detailed sketch of this approach.

[5] C.f. Exercise [FH91, 4.19], which can be solved by showing that the representation $\operatorname{Sym}^2(V_{n+1})$ is equivalent to the representation $U_{(n-2,2)}$, defined on [FH91, P. 54]. This can be done by explicitly computing the character of $\operatorname{Sym}^2(V_{n+1})$ (see [FH91, Chapter 2]) and $U_{(n-2,2)}$ (see [FH91, Eq. 4.33]). A decomposition of $U_{(n-2,2)}$ into irreps is given in the last displayed equation of [FH91, P. 57].

[6] This is the irreducible representation corresponding to the partition $((n+1)-2,2)$ of $n+1$ of $\mathfrak{S}_{n+1}$, a Specht module.

6.4.3 A_n

We will begin with (i). It is well known that $W(A_{n+1}) \cong \mathfrak{S}_{n+1}$ and we can explicitly describe the action of $W(A_n)$ in terms of the action of $\mathfrak{S}_{n+1}$ by permutation matrices via the identification $\mathcal{S}^n \cong \mathrm{Sym}^2(V_{n+1})$. For this, we explicitly write

$$\mathrm{Sym}^2(V_{n+1}) = \{A \in \mathcal{S}^{n+1} \,:\, Ae = 0\},$$

where, again, e is the all-ones vector. This can be done by identifying the root projectors $xx^\top$ with $x \in A_n$ with the projectors $M(e_i - e_j)$ with $e_i - e_j \in \mathbb{R}^{n+1}$. Let $\mathfrak{S}_{n+1}$ be the symmetric group on $n+1$ symbols. Define a group action of $\mathfrak{S}_{n+1}$ on $\mathbb{R}^{n+1}$ via

$$\mathfrak{S}_n \times \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}, \quad \sigma(v) := (v_{\sigma(1)}, \ldots, v_{\sigma(n+1)}). \tag{6.22}$$

For a Weyl group generator $S = I_{n+1} - aa^\mathsf{T}$ with $a = e_i - e_j$ one can straightforwardly verify that S is a permutation matrix that swaps the entries v_i and v_j:

$$Sv = \sigma(v), \quad \sigma = (i\ j).$$

As $\mathfrak{S}_n$ is generated by 2-cycles, it follows that $W(A_n)$ is a matrix representation (by permutation matrices) of $\mathfrak{S}_{n+1}$ acting on $\mathrm{Sym}^2(V_{n+1})$.

This identification enables us to use decomposition (6.21) and at this point we, in principle, have already found the decomposition proposed in the theorem. Clearly $U \cong \mathrm{span}\{I_n\}$. Furthermore, below we will show that $U_1(A_n), U_2(A_n)$, as given in the theorem, are indeed subrepresentations of $W(A_n) \cong \mathfrak{S}_{n+1}$ orthogonal to each other and $\mathrm{span}\{I_n\} \cong U$. We now proceed by comparing dimensions of the remaining summands. By the hook length formula [FH91, 4.12] we find $\dim(V_{n+1}) = n$ and $\dim(V_{((n+1)-2,2)}) = (n+1)(n-2)/2$. In Lemma 6.4.4 we will show that $\dim(U_2(A_n)) = n = \dim(V_{n+1})$, it then follows that $U_2(A_n) \cong V_{n+1}$. This also implies that $U_1(A_n) \cong V_{(n-2,2)}$, as the orthogonality of $U_1(A_n)$ and $U_2(A_n)$ implies that $U_1(A_n)$ is a subrepresentation of $V_{(n-2,2)}$, which, by the irreducibility of the latter, implies equivalence.

Therefore the following list of equivalences of representations is valid

$$U \cong \mathrm{span}\{I_n\},\ V_{n+1} \cong U_2(A_n),\ V_{(n-2,2)} \cong U_1(A_n),$$

which then, since U, V_{n+1}, and $V_{((n+1)-2,2)}$ are irreducible, finishes the proof of part (i) of the theorem.

We will conclude this part of the proof by showing that $U_1(A_n), U_2(A_n)$ are indeed subrepresentations of $W(A_n)$, are orthogonal to each other and computing $\dim(U_2(A_n)) = n$ as used above.

We first show orthogonality. It is straightforward to check that all operators in $U_i(R)$ for $R \in \{D_n, A_n\}$ and $i = 1, 2$ are traceless, so $\mathrm{span}\{I_n\} \perp U_i(R)$.

For $U_1(A_n) \perp U_2(A_n)$, we need to check that for orthogonal roots $x, y \in A_n$

$$0 = \langle P_i, M(x,y)\rangle = 2 \sum_{j \in \{1,\dots,n+1\}\setminus\{i\}} (x \cdot (e_i - e_j))(y \cdot (e_i - e_j)). \quad (6.23)$$

Every summand of the right hand side of (6.23) is zero, if $x = e_k - e_l$ and $y = e_s - e_t$ for $k, l, s, t \neq i$. Otherwise, if $x = \pm(e_i - e_k)$ and $y = e_s - e_t$, then $(x \cdot (e_i - e_j))(y \cdot (e_i - e_j))$ is only non-zero, if $j = s$ or $j = t$.

Then we get

$$(\pm(e_i - e_k) \cdot (e_i - e_s))((e_s - e_t) \cdot (e_i - e_s)) = \mp 1$$

and

$$(\pm(e_i - e_k) \cdot (e_i - e_t))((e_s - e_t) \cdot (e_i - e_t)) = \pm 1.$$

Thus, the sum of the right hand side of (6.23) is zero, which implies that the inner product is zero. Hence, all spaces in (i) are orthogonal.

Next, we show that the spaces are invariant under the action of the Weyl group. If x, y are orthogonal roots, then for $S \in W$ the roots Sx, Sy are orthogonal as well, because the Weyl group preserves orthogonality. This directly implies the invariance of $U_1(A_n)$. For the

invariance of $U_2(A_n)$ it suffices to observe that

$$S_{e_i-e_j} P_k (S_{e_i-e_j})^\top = \begin{cases} P_j, & \text{if } i = k \\ P_i, & \text{if } j = k \\ P_k, & \text{otherwise.} \end{cases}$$

As a last step we compute the dimension of the space $U_2(A_n)$.

Lemma 6.4.4. *For $n \geq 2$ it holds that* $\dim U_2(A_n) = n$.

Proof. By summing the generators P_i of $U_2(A_n)$ we obtain

$$\sum_{i=1}^{n+1} P_i = \sum_{x \in A_n} x x^\mathsf{T} - 2(n+1) I_n,$$

because each root projector xx^T, with $x \in A_n$, occurs in exactly two operators P_i and the roots x and $-x$ correspond to the same projector $xx^\mathsf{T} = (-x)(-x)^\mathsf{T}$. Since irreducible root systems are spherical 2-designs, (6.1) implies that

$$\sum_{x \in R} x x^\mathsf{T} = 2h I_n = 2(n+1) I_n.$$

Hence, $\sum_{i=1}^{n+1} P_i = 0$, and so the matrices P_i are linearly dependent. We now show that the matrices $P_1, \ldots, P_n$ are linearly independent, implying $\dim U_2(A_n) = n$. Suppose we have $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ with

$$\sum_{i=1}^{n} \lambda_i P_i = 0.$$

Let $\lambda = \lambda_1 + \cdots + \lambda_n$. We can write this equation as

$$\sum_{i=1}^{n} \sum_{j \in \{1,\ldots,n+1\} \setminus \{i\}} \lambda_i M(e_i - e_j) + 2\lambda I_n = 0.$$

For $i \neq j$, the projector $M(e_i - e_j)$ appears as a summand in P_i and $M(e_j - e_i)$ in P_j. Because $M(e_i - e_j) = M(e_j - e_i)$, rearranging the terms yields

$$\sum_{1 \leq i < j \leq n} (\lambda_i + \lambda_j) M(e_i - e_j) + \sum_{j=1}^{n} \lambda_j M(e_j - e_{n+1}) + 2\lambda I_n = 0.$$

As by (6.1),

$$I_n = \frac{1}{2(n+1)} \sum_{x \in R} xx^\mathsf{T} = \frac{1}{2(n+1)} \sum_{\substack{i,j=1,\dots,n+1 \\ i \neq j}} M(e_i - e_j)$$
$$= \frac{1}{n+1} \sum_{1 \leq i < j \leq n+1} M(e_i - e_j),$$

this becomes

$$\sum_{\substack{i,j=1,\dots,n \\ i \neq j}} \left(\lambda_i + \lambda_j + \frac{2\lambda}{n+1} \right) M(e_i - e_j) + \sum_{j=1}^{n} \left(\lambda_j + \frac{2\lambda}{n+1} \right) M(e_j - e_{n+1}) = 0. \tag{6.24}$$

Because the root projectors $\{M(e_i - e_j) : 1 \leq i < j \leq n+1\}$ are linearly independent[7], (6.24) implies that

$$\lambda_i + \lambda_j + \frac{2\lambda}{n+1} = 0, \quad 1 \leq i \neq j \leq n,$$
$$\text{and} \qquad \lambda_j + \frac{2\lambda}{n+1} = 0, \quad j = 1, \dots, n.$$

By subtracting the equations, it follows that $\lambda_1 = \dots = \lambda_n = 0$. □

6.4.4 D_n

We will proceed with (ii). The overall strategy is the same as in the A_n case. On the abstract level we consider the representation

$$\mathcal{S}^n \cong \mathrm{Sym}^2(\mathbb{R}^n) = \mathrm{Sym}^2(U + V_n).$$

We first obtain a decomposition of $\mathrm{Sym}^2(U + V_n)$ with respect to the action of the subgroup $\mathfrak{S}_n < W(D_n)$.

To this end, we first note that

$$\mathrm{Sym}^2(U + V_n) \cong \bigoplus_{a,b:\ a+b=2} \mathrm{Sym}^a(U) \otimes \mathrm{Sym}^b(V_n) \cong U \oplus V_n \oplus \mathrm{Sym}^2(V_n)$$

[7] This also follows from the fact that the root lattice is perfect and the number of root projectors coincides with the dimension of $\mathcal{S}^n$.

and the latter decomposes by (6.21), thus

$$\mathrm{Sym}^2(U + V_n) \cong U \oplus U \oplus V_n \oplus V_n \oplus V_{(n-2,2)}$$

as $\mathfrak{S}_n$-representations.

Now we examine how these (irreducible) $\mathfrak{S}_n$-subrepresentations behave under the action of $W(D_n)$, by directly comparing them to the modules given in the theorem.

First we show that the spaces in (ii) are indeed representations of $W(D_n)$. It is obvious that the spaces in (ii) are orthogonal. To verify that the spaces are indeed subrepresentations, note that for S_α for $\alpha^- = e_i - e_j$, $\alpha^+ = e_i + e_j$ and $\sigma = (i\ j) \in \Sigma_n$ we have

$$\begin{aligned} S_{\alpha^-} M(e_k, e_\ell) S_{\alpha^-}^\top &= M(e_{\sigma(k)}, e_{\sigma(\ell)}), \\ S_{\alpha^+} M(e_k, e_\ell) S_{\alpha^+}^\top &= M((-1)^{\delta_{k\in\{i,j\}}} e_{\sigma(k)}, (-1)^{\delta_{\ell\in\{i,j\}}} e_{\sigma(\ell)}), \end{aligned}$$

implying that $W(D_n)$ preserves I_n and maps the off-diagonal, respectively diagonal entries of a matrix to its off-diagonal, respectively diagonal entries. Hence, the spaces $U_1(D_n)$, $U_2(D_n)$ and $\mathrm{span}\{I_n\}$ are invariant under $W(D_n)$. The special case D_4 where $U_1(D_4)$ decomposes further into two 3-dimensional invariant subspaces will be treated at the end of this section. Now as $\mathfrak{S}_n$-representations we get (i.e. by comparing dimensions)

$$\mathrm{span}\{I_n\} \cong U,\ U_2(D_n) \cong V_n,$$

and, since they are already irreducible with respect to $\mathfrak{S}_n$, that these are irreducible $W(D_n)$-subrepresentations. Furthermore, by orthogonality, this implies

$$U_1(D_n) \cong U \oplus V_n \oplus V_{(n-2,2)}.$$

We are left to show that $U_1(D_n)$ is irreducible for $n \geq 5$ and to obtain a decomposition into irreducible subrepresentations for $n = 4$.

It is easy to see that, with respect to the action of $\mathfrak{S}_n$,

$$\begin{aligned} U \oplus V_n \cong L &:= \mathrm{span}\left\{ M\Big(e_i, \sum_{j\in\{1,\dots,n\}\setminus\{i\}} e_j\Big) \,:\, i = 1, \dots, n \right\} \\ &= \left\{ \sum_{1\leq i<j\leq n} (a_i + a_j) M(e_i, e_j) \,:\, a_1, \dots, a_n \in \mathbb{R} \right\} \subset U_2(D_n) \end{aligned}$$

and

$$U \cong L_1 \coloneqq \operatorname{span}\left\{ \sum_{1\le i<j\le n} M(e_i, e_j) \right\}.$$

Hence, by orthogonality of U and V_n,

$$V_n \cong L_1^\perp \coloneqq \left\{ \sum_{1\le i<j\le n} (a_i + a_j) M(e_i, e_j) \;:\; a_1, \dots, a_n \in \mathbb{R},\, \sum_{i=1}^{n} a_i = 0 \right\}.$$

If $U_2(D_n)$ is not an irreducible $W(D_n)$-representation, then, by Maschke's theorem, either $L_1, L_1^\perp$ or $L = L_1 \perp L_1^\perp$ is an irreducible $W(D_n)$-representation.

We can directly see that L_1 and L are not even $W(D_n)$-invariant: considering the action of the element $\alpha = e_1 + e_2 \in W(D_n)$ gives

$$S_\alpha \Big(\underbrace{\sum_{1\le i<j\le n} M(e_i, e_j)}_{\in L_1 \subset L} \Big) S_\alpha^\top = M(e_1, e_2) - \sum_{i\in\{1,2\}} \sum_{k=3}^{n} M(e_i, e_k) + \sum_{3\le i,j\le n} M(e_i, e_j) \notin L,$$

by showing that a certain system of linear equations has no solutions.

We are left with the case of $L_1^\perp$ to consider. Here we fix the element

$$X \coloneqq M(e_1, \sum_{j\in\{2,\dots,n\}} e_j) - M(e_2, \sum_{j\in\{1,\dots,n\}\setminus\{2\}} e_j) \in L_1^\perp.$$

Now, choosing $\alpha = e_3 + e_4$, we can show that $S_\alpha X S_\alpha \notin L_1 \oplus L_1^\perp$ for $n \ge 5$, again by considering a system of linear equations.

However, if $n = 4$, the system allows for a solution and the space $L_1^\perp$ can be written as

$$L_1^\perp = \left\{ \begin{pmatrix} 0 & a & b & -c \\ a & 0 & c & -b \\ b & c & 0 & -a \\ -c & -b & -a & 0 \end{pmatrix} \;:\; a, b, c \in \mathbb{R} \right\},$$

which can be shown to be invariant under $W(D_4)$. Thus, $L_1^\perp$ is irreducible and $U_2(D_4)$ splits into two $W(D_4)$-irreducible subspaces as

$$U_2(D_4) = L_1^\perp \oplus L_2, \quad L_2 = \left\{ \begin{pmatrix} 0 & a & b & c \\ a & 0 & c & b \\ b & c & 0 & a \\ c & b & a & 0 \end{pmatrix} \;:\; a, b, c \in \mathbb{R} \right\}$$

with $L_2 \cong U \oplus V_{(n-2,2)}$.

It remains to prove that the irreducible subspaces for the special case D_4 are inequivalent, despite having the same dimension.

We will do this by showing a more general statement, that is, if T is an intertwiner with respect to the action of $W(D_n)$, then $M(x,y)$ is an eigenvector of T for orthogonal roots $x, y \in D_n$.

In the case of D_4, all three subspaces $U_2(D_n)$, $L_1^\perp$ and L_2 contain an operator $M(x,y)$ for orthogonal roots $x, y \in D_4$. This shows in particular that the intertwiner T is either identically zero on one of the three subspaces or $U_2(D_n)$, $L_1^\perp$ and L_2 or T must preserve the three subspaces. By Schur's lemma, this implies that they are inequivalent.

To see this, note that for $\sigma(x) \in W$ it holds that

$$\sigma(x)M(x,y)\sigma(x)^\mathsf{T} = \sigma(y)M(x,y)\sigma(y)^\mathsf{T} = -M(x,y),$$

so $M(x,y)$ is contained in the subspace

$$U_{xy} := \{X \in S^n \,:\, \sigma(x)X\sigma(x)^\mathsf{T} = \sigma(y)^\mathsf{T}X\sigma(y)^\mathsf{T} = -X\}.$$

Let $X \in U_{xy}$. Since T commutes with the action of W, it follows

$$\sigma(x)T(X)\sigma(x)^\mathsf{T} = T(\sigma(x)X\sigma(x)^\mathsf{T}) = -T(X) = \sigma(y)X\sigma(y)^\mathsf{T},$$

hence $T(U_{xy}) \subseteq U_{xy}$. Now, consider the $M(x,y)$ with $x = e_1 + e_3$ and $y = e_3 + e_4$ and assume that $X = \sum_{1\le i\le j\le n} c_{ij}M(e_i,e_j) \in U_{xy}$. Due to

$$\sigma(x)e_i = \begin{cases} -e_i, & \text{if } i = 1,2 \\ e_i, & \text{otherwise} \end{cases} \qquad \sigma(y)e_i = \begin{cases} -e_i, & \text{if } i = 3,4 \\ e_i, & \text{otherwise} \end{cases}$$

it follows that

$$\begin{aligned} -X = \sigma(x)X\sigma(x)^\top =& M(e_1,e_2) + M(e_1,e_1) + M(e_2,e_2) \\ &- \sum_{i>2} c_{2i}M(e_1,e_i) + c_{1i}M(e_2,e_i) + \sum_{i,j>2} c_{ij}M(e_i,e_j), \end{aligned}$$

hence $c_{1i} = c_{2i}$ and $c_{ij} = 0$ for all other cases. Acting with $\sigma(y)$ on X yields

$$\begin{aligned} -X &= \sigma(y)\Big(\sum_{i>2} c_{1i}(M(e_1,e_i) + M(e_2,e_i))\Big)\sigma(y)^\top \\ &= -c_{14}M(e_1,e_3) - c_{13}M(e_1,e_4) - c_{14}M(e_2,e_3) - c_{13}M(e_2,e_4) + \sum_{i>5} c_{1i}M(e_1,e_i), \end{aligned}$$

so $c_{14} = c_{13}$ and $c_{1i} = 0$ for $i \neq 3, 4$. Hence, $X = cM(x, y)$ for some constant $c \in \mathbb{R}$ and U_{xy} is one-dimensional. As $T(U_{xy}) \subseteq U_{xy}$, this shows that $M(x, y)$ is an eigenvector of the intertwiner T. The argument for general orthogonal roots $x, y \in D_n$ follows in the same manner.

6.4.5 E_n

To give an elementary proof that $\mathcal{T}_0^n$ is irreducible with respect to the action of $W(E_n)$, we will use (ii) of Theorem 6.4.2.

In all three cases we consider the embedding of the root system E_n into $\mathbb{R}^8$, as defined in Section 6.2.3. For $n \in \{6, 7\}$ the space $\mathcal{T}_0^n$ embeds into $\operatorname{span}\{xx^\mathsf{T} : x \in E_n\} \subset \mathcal{S}^8$ via

$$\mathcal{T}_0^n \cong \begin{cases} \{X \in \mathcal{T}_0^8 : X(e_7 - e_8) = 0,\, X(e_6 - e_7) = 0\}, & \text{if } n = 6 \\ \{X \in \mathcal{T}_0^8 : X(e_7 - e_8) = 0\}, & \text{if } n = 7. \end{cases}$$

Further, we embed the root systems D_n for $n \leq 8$ into $\mathbb{R}^8$ by adding zero coordinates to the roots. Let D_{s_n} be the largest root system of type D that is contained in E_n, that is $D_{s_6} = D_5$, $D_{s_7} = D_6$ and $D_{s_8} = D_8$. Since $W(D_{s_n})$ is a subgroup of the Weyl group $W(E_n)$, Schur's lemma implies that every intertwiner T with respect to $W(E_n)$ is a scalar multiple of the identity on $U_i(D_{s_n})$.The intertwiner commutes with the group action, thus, it is also a scalar multiple on

$$W(E_n) \cdot U_i(D_{s_n}) := \{SXS^\top : X \in U_i(D_{s_n})\},$$

so $W(E_n) \cdot U_i(D_{s_n})$ is an irreducible subspace for the action of $W(E_n)$. Hence, to prove the irreducibility of $\mathcal{T}_0^n$, it suffices to prove that

$$\mathcal{T}_0^n = W(E_n) \cdot U_2(D_{s_s}). \tag{6.25}$$

First, we show that the two orbits $W(E_n) \cdot U_i(D_{s_s})$ collapse to one subspaces under the action of $W(E_n)$:

Lemma 6.4.5. *It holds that $U_1(D_{s_n}) \subset W(E_n) \cdot U_2(D_{s_n})$ and in particular,*

$$T_0^{s_n} \cong \operatorname{span}\{U_1(D_{s_n}), U_2(D_{s_n})\} \subset W(E_n) \cdot U_2(D_{s_n}),$$

where the first equivalence is a consequence of Theorem 6.4.2 (ii).

The lemma already shows the identity (6.25) for $n = 8$ and therefore the irreducibility of $\mathcal{T}_0^8$ with respect to the action of $W(E_8)$.

Proof. It suffices to show that for $M(e_1 + e_2, e_3 - e_4) \in U_1(D_{s_n})$ and $M(e_1 + e_2, e_1 - e_2) \in U_2(D_{s_n})$ it holds that

$$M(e_1 + e_2, e_3 - e_4) \in W(E_n) \cdot M(e_1 - e_2, e_1 + e_2).$$

We have

$$e_1 - e_2 = (1, -1, 0, 0, 0, 0, 0, 0) \quad \overset{\sigma(x_1)}{\longmapsto} \quad \frac{1}{2}(1, -1, -1, 1, 1, 1, 1, 1) =: y$$
$$\text{for} \quad x_1 = \frac{1}{2}(1, -1, 1, -1, -1, -1, -1, -1) \in E_n.$$

Moreover,

$$y = \frac{1}{2}(1, -1, -1, 1, 1, 1, 1, 1) \quad \overset{\sigma(x_2)}{\longmapsto} \quad (0, 0, -1, 1, 0, 0, 0, 0) = -(e_3 - e_4)$$
$$\text{for} \quad x_2 = \frac{1}{2}(1, -1, 1, -1, 1, 1, 1, 1) \in E_n.$$

Since both $\sigma(x_1)$ and $\sigma(x_2)$ stabilize $e_1 + e_2$, it follows that

$$\sigma(x_2)\sigma(x_1)M(e_1+e_2, e_1-e_2)\sigma(x_1)^{\mathsf{T}}\sigma(x_2)^{\mathsf{T}} = -M(e_1+e_2, e_3-e_4). \quad \square$$

It remains to prove (6.25) for $n \in \{6, 7\}$.

Proposition 6.4.6. *We have*

$$\dim W(E_n) \cdot U_2(D_{s_n}) \geq \dim \mathcal{T}_0^{s_n} + n = \dim \mathcal{T}_0^n,$$

so $W(E_n) \cdot U_2(D_{s_n}) \cong \mathcal{T}_0^n$.

Proof. We identify $\mathcal{T}_0^{s_n}$ with the space of all traceless symmetric matrices whose last $n - s_n$ rows respectively columns are zeros. As a consequence of Lemma 6.4.5, $\mathcal{T}_0^{s_n} \subset W(E_n) \cdot U_2(D_{s_n})$, so it suffices to find n matrices $X_1, \ldots, X_n \in (W(E_n) \cdot M(x, y)) \setminus \mathcal{T}_0^{s_n}$ such that

$$\dim \operatorname{span}\{\mathcal{T}_0^{s_n}, X_1, \ldots, X_n\} = \dim \mathcal{T}_0^{s_n} + n. \tag{6.26}$$

Therefore, observe that for each root $z \in E_n \setminus D_{s_n}$ we can find a tuple of roots $x, y \in D_{s_n}$ and an element $S \in W(E_n)$ such that

$$Sx = z \quad \text{and} \quad Sy = y. \tag{6.27}$$

The action of S maps $xx^\mathsf{T} - yy^\mathsf{T} \in \mathcal{T}_0^{s_n}$ to $zz^\mathsf{T} - yy^\mathsf{T} \notin \mathcal{T}_0^{s_n}$. To see (6.27), if $z = \frac{1}{2}(a_1, \ldots, a_8) \in E_n \setminus D_{s_n}$ with $a_i \in \{\pm 1\}$, choose

$$x = (a_1, a_2, 0, \ldots, 0), \quad y = (a_1, -a_2, 0, \ldots, 0) \quad \text{and}$$
$$S = \sigma(z') \quad \text{with} \quad z' = \frac{1}{2}(a_1, a_2, -a_3, \ldots, -a_8).$$

Then, one can directly verify that $Sx = z$ and $Sy = y$.

Now, choose a set of linearly independent roots $z_1, \ldots, z_n \in E_n \setminus D_{s_n}$. Such a set exists, for example, take the roots $z_1, \ldots, z_n \in E_n \setminus D_{s_n}$ such that the i-th and $(i+1)$-th entry of root z_i for $1 \leq i \leq n-1$ are negative and the remaining entries positive, and for z_n we set the first and the n-th entry to be negative and the remaining ones positive.

Additionally, choose $y_1, \ldots, y_n \in D_{s_n}$. Then the matrices $X_i = z_i z_i^\mathsf{T} - y_i y_i^\mathsf{T}$ lie in $W(E_n) \cdot U_2(D_{s_n}) \setminus \mathcal{T}_0^{s_n}$. These matrices are linearly independent since the last row of $z_i z_i^\mathsf{T} - y_i y_i^\mathsf{T}$ is given by the vector $\pm 1/2 z_i$ and vectors z_i were chosen to be linearly independent. Since the last row of every matrix in $\mathcal{T}_0^{s_n}$ consists of only zeros, it follows that adding these vectors to $\mathcal{T}_0^{s_n}$ increases the dimension of their joint span by n, which proves (6.26). □

Proof of Theorem 6.4.1

To prove Theorem 6.4.1 it remains to compute

$$Q[A] = \lambda \operatorname{Tr} A^2$$

for A contained in one of the spaces given in Theorem 6.4.2.

We first evaluate Q at the identity matrix. We have

$$Q[I_n] = \sum_{r \in R} (r^\mathsf{T} r)^2 = 4|R|,$$

and using $\operatorname{Tr} I_n = n$ we see that $\lambda = 4h$, where h is the Coxeter number of the root system R.

Note that for $R = A_n$ or $R = D_n$ we can find $x, y \in R$ with $x \cdot y = 0$ and $\{x, y\} \neq \{e_i - e_j, e_i + e_j\}$ such that $M(x, y) \in U_1(R)$. In the case of D_4 we can find such an element $M(x, y)$ in both of the two

irreducible subspaces decomposing $U_1(D_4)$. Then

$$\begin{aligned} Q[M(x,y)] &= \sum_{r\in R} M(x,y)[r]^2 \\ &= 4\sum_{r\in R} (x\cdot r)^2(y\cdot r)^2. \end{aligned}$$

We only have to consider roots r, with $r\cdot x \neq 0$ and $r\cdot y \neq 0$, which implies $(r\cdot x)^2 = (r\cdot y)^2 = 1$. For $R = A_n$ we can find 8 roots fulfilling this condition, for $R = D_n$ there are 16. Hence,

$$Q[M(x,y)] = \begin{cases} 32 & \text{for } R = A_n, \\ 64 & \text{for } R = D_n. \end{cases}$$

For the matrices $M(e_i - e_j, e_i + e_j) \in U_2(D_n)$, the result is similarly

$$Q[M(e_i - e_j, e_i + e_j)] = 4\sum_{r\in D_n} ((e_i + e_j)\cdot r)^2((e_i - e_j)\cdot r)^2.$$

If $r = \pm e_i \pm e_j$, the summand is zero. Otherwise, if $(r\cdot(e_i + e_j))^2 = 1$, it follows $(r\cdot(e_i - e_j))^2 = 1$, and there are exactly $8(n-2)$ such roots $r \in D_n$. Hence, $Q[M(e_i - e_j, e_i + e_j)] = 32(n-2)$.

In all three cases, the normalizing factor is

$$\operatorname{Tr} M(x,y)^2 = 2(x\cdot x)(y\cdot y) + 2(x\cdot y)^2 = 8.$$

So we obtain eigenvalues 4 on $U_1(A_n)$, respectively 8 and $4(n-2)$ on $U_1(D_n)$ and $U_2(D_n)$.

For $R = A_n$ we have to compute the eigenvalue for $U_2(A_n)$, so we may evaluate $Q(P_1)$. Observe that

$$P_1[r]^2 = \left(\sum_{j\in\{2,\dots,n+1\}} ((e_1 - e_j)\cdot r)^2 - 4\right)^2.$$

If $r = \pm(e_1 - e_j)$ for some $j \in \{2,\dots,n+1\}$, then we get $(r,r)^2 = 4$ and $((e_1 - e_j)\cdot r)^2 = 1$ for all other j. This amounts to

$$P_1[r]^2 = (4 + (n-1) - 4)^2 = (n-1)^2.$$

If $r = (e_k - e_l)$ with $k, l \neq 1$, it follows $(r\cdot(e_1 - e_k))^2 = (r\cdot(e_1 - e_l))^2 = 1$ and all other summands are zero. So we get

$$P_1[r]^2 = (2-4)^2 = 4.$$

There are $2n$ roots of type $\pm(e_1 - e_j)$ and accordingly $n(n-1)$ of type $(e_k - e_l)$ with $k, l \neq 1$. This results in

$$Q[P_1] = 2n(n-1)^2 + 4n(n-1) = 2n(n-1)(n+1).$$

Now we compute $\operatorname{Tr} P_1^2 = \langle P_1, P_1 \rangle$ and get

$$\begin{aligned}\langle P_1, P_1 \rangle &= \Big\langle \sum_{j\in\{2,\dots,n+1\}} M(e_1 - e_j) - 2I_n, \sum_{j\in\{2,\dots,n+1\}} M(e_1 - e_j) - 2I_n \Big\rangle \\ &= \sum_{j,k\in\{2,\dots,n+1\}} \langle M(e_1 - e_j), M(e_1 - e_k)\rangle - 4 \sum_{j\in\{2,\dots,n+1\}} \langle M(e_1 - e_j), I_n\rangle + 4n \\ &= \sum_{j,k\in\{2,\dots,n+1\}} ((e_1 - e_k)\cdot(e_1 - e_l))^2 - 4n.\end{aligned}$$

The first sum equals

$$\sum_{j\in\{2,\dots,n+1\}} ((e_1 - e_j)\cdot(e_1 - e_j))^2 + \sum_{2\leq j\neq k\leq n+1} ((e_1 - e_j)\cdot(e_1 - e_k))^2 = 4n + n(n-1).$$

Hence, together we have $\langle P_1, P_1\rangle = n(n-1)$ and the eigenvalue associated with the eigenspace $U_2(A_n)$ is $2(n+1) = 2h$.

The remaining eigenvalues for E_6, E_7, E_8 are given by the fact that these root systems form spherical 4-designs. Then, by (6.11), $Q[H] = \frac{8h}{n+2} \operatorname{Tr} H^2$. □

6.4.6 Orthogonal sum of irreducible root systems

In this section, we want to compute the eigenvalues of the quadratic form Q on the orthogonal sum of irreducible root systems $R = R_1 \perp \ldots \perp R_m$. For this we write

$$Q_{R_i}[H] = \sum_{x\in R_i} H[x]^2$$

to distinguish between the quadratic form on different root systems R_i. Let n_i be the rank of R_i. Furthermore, let $n = n_1 + \cdots + n_m$. Write each $x \in \mathbb{R}^n$ as $(x_1, \dots, x_m)$ with $x_i \in \mathbb{R}^{n_i}$ and every root $r \in R$ as $(0, \dots, 0, r_i, 0, \dots, 0)$ with $r_i \in R_i$ and $0 \in \mathbb{R}^{n_j}$ accordingly. To compute the eigenvalues of Q_R on $\mathcal{S}^n$, we identify $\mathcal{S}^n$ in a similar fashion: Each $H \in \mathcal{S}^n$ can be seen as a vector of block matrices

$$H \cong (H_{1,1}, \dots, H_{m,m}, H_{1,2}, \dots, H_{m-1,m}) \iff H = \begin{pmatrix} H_{1,1} & H_{1,2} & \cdots & H_{1,m} \\ H_{1,2}^{\mathsf{T}} & H_{2,2} & \cdots & H_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ H_{1,m}^{\mathsf{T}} & H_{2,m}^{\mathsf{T}} & \cdots & H_{m,m} \end{pmatrix}, \tag{6.28}$$

where $H_{i,i} \in \mathcal{S}^{n_i}$ and $H_{i,j} \in \mathbb{R}^{n_i \times n_j}$ for $i \neq j$. This way, we identify

$$\mathcal{S}^n \cong \mathcal{S}^{n_1} \perp \ldots \perp \mathcal{S}^{n_m} \perp\!\!\!\perp_{1 \leq i < j \leq m} \mathbb{R}^{n_i \times n_j}. \tag{6.29}$$

Furthermore, let $\mathcal{D}$ be the m-dimensional space that is spanned by the diagonal matrices

$$(I_{n_1}, 0, \dots, 0), (0, I_{n_2}, 0, \dots, 0), \dots, (0, \dots, 0, I_{n_m}, 0, \dots, 0).$$

We are particularly interested in the case where each component of the root system R has the same Coxeter number. In this case R is of the form

$$R = (A_{n_a})^{m_a} \perp (D_{n_d})^{m_d} \perp (E_{n_e})^{m_e}, \tag{6.30}$$

where $(A_{n_a})^{m_a}$, $(D_{n_d})^{m_d}$ respectively $(E_{n_e})^{m_e}$ are orthogonal sums of m_a, m_d respectively m_e irreducible roots systems A_{n_a}, D_{n_d} respectively E_{n_e}, and $m = m_a + m_d + m_e$, $n = m_a n_a + m_d n_d + m_e n_e$.

Theorem 6.4.7. *Let $R = \perp_{i=1}^{m} R_i$ be the orthogonal sum of irreducible root systems $R_i \in \{A_{n_i}, D_{n_i}, E_{n_i}\}$, where n_i is the rank of R_i. We identify $\mathcal{S}^n$ as in* (6.29).

(i) We have

$$Q_R[H] = Q_{R_1}[H_{1,1}] + \cdots + Q_{R_m}[H_{m,m}], \tag{6.31}$$

so the quadratic form only depends on the diagonal entries $H_{i,i} \in \mathcal{S}^{n_i}$ and the eigenvalues of Q_R are the eigenvalues of all Q_{R_i} and additionally the eigenvalue 0 with multiplicity $\sum_{1 \leq i < j \leq m} n_i n_j$.

(ii) If each component root system has the same Coxeter number h, we can write R as in (6.30). *The space of traceless matrices $\mathcal{T}_0^n$*

then decomposes into eigenspaces of Q_R:

$$\begin{aligned}\mathcal{T}_0^n \;=\;& U_1(A_{n_a})^{m_a} \perp U_2(A_{n_a})^{m_a} \\ &\perp U_1(D_{n_d})^{m_d} \perp U_2(D_{n_d})^{m_d} \\ &\perp (\mathcal{T}_0^{n_e})^{m_e} \\ &\perp \mathcal{D} \cap \mathcal{T}_0^n,\end{aligned} \tag{6.32}$$

where the exponents refer to the direct sum of the eigenspaces of $Q_{A_{n_a}}, Q_{D_{n_d}}$ and $Q_{E_{n_e}}$. The eigenspace $\mathcal{D} \cap \mathcal{T}_0^n$ belongs to the eigenvalue $4h$ and has dimension $m-1$.

Remark 6.4.8. *The decomposition* (6.32) *does not change when D_4 is considered because the quadratic form has the same eigenvalues on both irreducible subspaces that decompose $U_1(D_4)$.*

Proof. (i) Let $H \in \mathcal{S}^n$. We write H as in (6.28). For a root $r = (0, \dots, 0, r_i, 0, \dots, 0) \in R$ it follows

$$H[r] = H_{i,i}[r_i],$$

so $H[r]$ does not depend of the off-diagonal entries $(0, \dots, 0, H_{i,j}, 0, \dots, 0)$ for $i \neq j$ of H.

Since every root in R is of this form, this directly implies (6.31). This also shows that the eigenvalues of Q_R coincide with the eigenvalues of Q_{R_i} with the same multiplicity. The only additional eigenvalue we get is 0, which is obtained by evaluating $Q_R[H]$ for matrices $H \in \mathcal{S}^n$, where all diagonal entries $H_{ii} = 0 \in \mathcal{S}^{n_i}$. Due to the identification (6.29), the space of these matrices has dimension $\sum_{1 \leq i < j \leq m} n_i n_j$, which gives the multiplicity of the eigenvalue 0.

(ii) If each component root system of R has the same Coxeter number, the space $\mathcal{D}$ is an eigenspace of Q_R because

$$\begin{aligned}Q_R\Big[(0, \dots, 0, I_{n_a}, 0 \dots, 0)\Big] &= Q_{A_{n_a}}[I_{n_a}] = 4h, \\ Q_R\Big[(0, \dots, 0, I_{n_d}, 0 \dots, 0)\Big] &= Q_{D_{n_d}}[I_{n_d}] = 4h, \\ Q_R\Big[(0, \dots, 0, I_{n_e}, 0 \dots, 0)\Big] &= Q_{E_{n_e}}[I_{n_e}] = 4h.\end{aligned}$$

Hence, $\mathcal{S}^n$ decomposes into eigenspaces of Q_R as

$$\mathcal{S}^n = U_1(A_{n_a})^{m_a} \perp U_2(A_{n_a})^{m_a} \perp U_1(D_{n_d})^{m_d} \perp U_2(D_{n_d})^{m_d} \perp (\mathcal{T}_0^{n_e})^{m_e} \perp \mathcal{D}.$$

All eigenspaces but $\mathcal{D}$ lie in the space $\mathcal{T}_0^n$, hence equation (6.32) holds. To see that $\mathcal{D} \cap \mathcal{T}_0^n$ has dimension $m-1$, note that it contains all diagonal matrices of the form

$$(c_1 I_{n_1}, \ldots, c_m I_{n_m}, 0, \ldots, 0) \quad \text{with} \quad c_1 n_1 + \cdots + c_m n_m = 0.$$

Since $\mathcal{D}$ has dimension m, it follows that $\mathcal{D} \cap \mathcal{T}_0^n$ has dimension $m-1$. □

6.5 Concrete results

6.5.1 Dimension 8

Mordell [Mor38] showed that the root lattice E_8 is the only even unimodular lattice in dimension 8. By [CKM+22] E_8 is universally optimal and unique among periodic point configurations. The fact that it is a local minimum for all Gaussian potential functions was established in [SS06]. Coulangeon [Cou06] used (6.12) to provide an alternative proof.

6.5.2 Dimension 16

Witt [Wit41] proved that there exist exactly two even unimodular lattices in dimension 16: D_{16}^+ and $E_8 \perp E_8$. Both lattices have the same theta series E_4^2, but their root systems differ as we have $D_{16}^+(2) = D_{16}$ and, respectively, $E_8 \perp E_8(2) = E_8 \perp E_8$. The eigenvalues of the quadratic form (6.13) are by Theorem 6.4.1 and Theorem 6.4.7

$$8\ (120\times), 56\ (15\times) \quad \text{respectively} \quad 0\ (64\times), 24\ (70\times), 120\ (1\times).$$

Therefore, by Corollary 6.3.2, D_{16}^+ is a local minimum for f_α-potential energy whenever α is large enough. By Corollary 6.3.2 the other lattice $E_8 \perp E_8$ is a saddle point whenever α is large enough. The following numerical computations strongly suggest that $E_8 \perp E_8$ is in fact a saddle point for all values of α.

Using SageMath [Tea] we arrive at the following plot for the eigenvalues of the Hessian of the function $L \mapsto \mathcal{E}(f_\alpha, L)$ at D_{16}^+ and at $E_8 \perp E_8$.

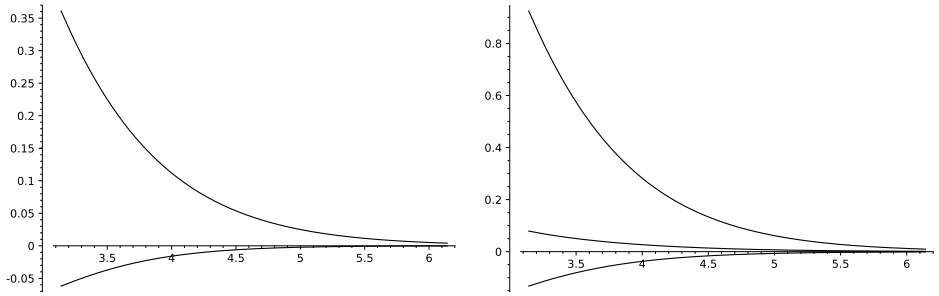


Figure 6.1: The eigenvalues of the Hessian for D_{16}^+ (two different eigenvalues, left) and $E_8 \perp E_8$ (three different eigenvalues, right) depending on the parameter α.

We introduce the following notation: The value in (6.15) we denote by $\mu(L, \lambda, \alpha)$. We consider $\alpha = \pi$, then

$$\begin{aligned} \mu(D_{16}^+, 8, \pi) &= -0.06196\ldots & \qquad \mu(E_8 \perp E_8, 0, \pi) &= -0.13245\ldots \\ \mu(D_{16}^+, 56, \pi) &= 0.36093\ldots & \qquad \mu(E_8 \perp E_8, 24, \pi) &= 0.07899\ldots \\ & & \qquad \mu(E_8 \perp E_8, 120, \pi) &= 0.92480\ldots \end{aligned}$$

We show in Section 6.5.4 how to translate numerical computations into rigorous bounds.

6.5.3 Dimension 24

The Niemeier lattices are the even unimodular lattices in dimension 24 which have vectors of squared norm 2. A classification of Niemeier gave that there are 23 Niemeier lattices and Venkov realized that they can be characterized by their root system. The theta series of a Niemeier lattice L with root system $L(2)$ is the modular form of weight 12

$$\Theta_L(\tau) = E_4^3(\tau) + (|L(2)| - 720)\Delta(\tau) = 1 + |L(2)|q + \cdots .$$

The cusp form of weight 16 is $E_4\Delta$. We apply Theorem 6.3.1 together with Theorem 6.4.1 and Theorem 6.4.7 to determine the signature of the Hessian at $\alpha = \pi$. We collect our results in Table 6.2. For large values of α Corollary 6.3.2 shows that only the Niemeier lattices with

irreducible root systems, namely A_{24} and D_{24}, are local minima for f_α-potential energy. All other Niemeier lattices are saddle points for f_α-potential energy for α large enough.

$L(2)$	$\lvert L(2)\rvert$	h	λ	multiplicity	$\mu(L,\lambda,\pi)$
A_1^{24}	48	2	0	276	$0.0018\ldots$
			8	23	$0.1044\ldots$
A_2^{12}	72	3	0	264	$-0.0050\ldots$
			6	24	$0.0718\ldots$
			12	11	$0.1488\ldots$
A_3^8	96	4	0	252	$-0.0120\ldots$
			4	16	$0.0392\ldots$
			8	24	$0.0905\ldots$
			16	7	$0.1931\ldots$
A_4^6	120	5	0	240	$-0.0189\ldots$
			4	30	$0.0323\ldots$
			10	24	$0.1092\ldots$
			20	5	$0.2375\ldots$
$A_5^4D_4$	144	6	0	230	$-0.0259\ldots$
			4	36	$0.0253\ldots$
			8	9	$0.0766\ldots$
			12	20	$0.1279\ldots$
			24	4	$0.2818\ldots$
D_4^6	144	6	0	240	$-0.0259\ldots$
			8	54	$0.0766\ldots$
			24	5	$0.2818\ldots$
A_6^4	168	7	0	216	$-0.0328\ldots$
			4	56	$0.0184\ldots$
			14	24	$0.1466\ldots$
			28	3	$0.3262\ldots$
$A_7^2D_5^2$	192	8	0	214	$-0.0398\ldots$
			4	40	$0.0114\ldots$
			8	20	$0.0627\ldots$
			12	8	$0.1140\ldots$
			16	14	$0.1653\ldots$
			32	3	$0.3705\ldots$
A_8^3	216	9	0	192	$-0.0467\ldots$
			4	81	$0.0045\ldots$
			18	24	$0.1840\ldots$
			36	2	$0.4149\ldots$

Table 6.2: The eigenvalues of the Hessian of the Niemeier lattices for $\alpha = \pi$.

$L(2)$	$\lvert L(2)\rvert$	h	λ	multiplicity	$\mu(L,\lambda,\pi)$
$A_9^2D_6$	240	10	0	189	$-0.0537\ldots$
			4	70	$-0.0024\ldots$
			8	15	$0.0488\ldots$
			16	5	$0.1514\ldots$
			20	18	$0.2027\ldots$
			40	2	$0.4592\ldots$
D_6^4	240	10	0	216	$-0.0537\ldots$
			8	60	$0.0488\ldots$
			16	20	$0.1514\ldots$
			40	3	$0.4592\ldots$
E_6^4	288	12	0	216	$-0.0676\ldots$
			12	80	$0.0862\ldots$
			48	3	$0.5479\ldots$
E_6^4	288	12	0	216	$-0.0676\ldots$
			12	80	$0.0862\ldots$
			48	3	$0.5479\ldots$
$A_{11}D_7E_6$	288	12	0	185	$-0.0676\ldots$
			4	54	$-0.0163\ldots$
			8	21	$0.0349\ldots$
			12	20	$0.0862\ldots$
			20	6	$0.1888\ldots$
			24	11	$0.2401\ldots$
			48	2	$0.5479\ldots$
A_{12}^2	312	13	0	144	$-0.0746\ldots$
			4	130	$-0.0233\ldots$
			26	24	$0.2588\ldots$
			52	1	$0.5923\ldots$
D_8^3	336	14	0	192	$-0.0815\ldots$
			8	84	$0.0210\ldots$
			24	21	$0.2262\ldots$
			56	2	$0.6366\ldots$
$A_{15}D_9$	384	16	0	135	$-0.0954\ldots$
			4	104	$-0.0441\ldots$
			8	36	$0.0071\ldots$
			28	8	$0.2636\ldots$
			32	15	$0.3149\ldots$
			64	1	$0.7253\ldots$

TABLE 6.2. (continued).

$L(2)$	$\lvert L(2)\rvert$	h	λ	multiplicity	$\mu(L,\lambda,\pi)$
$A_{17}E_7$	432	18	0	119	$-0.1093\ldots$
			4	135	$-0.0580\ldots$
			16	27	$0.0958\ldots$
			36	17	$0.3523\ldots$
			72	1	$0.8140\ldots$
$D_{10}E_7^2$	432	18	0	189	$-0.1093\ldots$
			8	45	$-0.0067\ldots$
			16	54	$0.0958\ldots$
			32	9	$0.3010\ldots$
			72	2	$0.8140\ldots$
D_{12}^2	528	22	0	144	$-0.1371\ldots$
			8	132	$-0.0345\ldots$
			40	22	$0.3758\ldots$
			88	1	$0.9914\ldots$
A_{24}	600	25	4	275	$-0.1067\ldots$
			50	24	$0.4832\ldots$
$D_{16}E_8$	720	30	0	128	$-0.1928\ldots$
			8	120	$-0.0902\ldots$
			24	35	$0.1150\ldots$
			56	15	$0.5254\ldots$
			120	1	$1.3462\ldots$
E_8^3	720	30	0	192	$-0.1928\ldots$
			24	105	$0.1150\ldots$
			120	2	$1.3462\ldots$
D_{24}	1104	46	8	276	$-0.2014\ldots$
			88	23	$0.8246\ldots$

TABLE 6.2. (continued).

6.5.4 Dimension 32

In dimension 32 the even unimodular lattice have not been classified yet. Some partial results are known: There are at least 80 million of them, see Serre [Ser73]. King [Kin03] showed that there are at least ten million even unimodular lattices without roots in dimension 32. Kervaire [Ker94] classified all indecomposable even unimodular lattices in dimension 32 that possess a full root system.

In general an even unimodular lattice need not even be a critical point for the Gaussian potential function. The first such examples can

be found in dimension 32, we briefly discuss one of them.

For example there exists a lattice $L \subseteq \mathbb{R}^{32}$ with complete root system $A_1^8 A_3^8$, see Kervaire [Ker94]. We split the summation in the gradient into the contribution of the root system and the contribution of all larger shells

$$\begin{aligned}\langle \nabla \mathcal{E}(f_\alpha, L), H\rangle &= -\alpha \sum_{x \in L\setminus\{0\}} H[x] e^{-\alpha\|x\|^2} \\ &= -\alpha e^{-2\alpha} \left(\sum_{x \in L(2)} H[x] \right) - \alpha \left(\sum_{x \in L\setminus(\{0\}\cup L(2))} H[x] e^{-\alpha\|x\|^2} \right).\end{aligned}$$

We firstly evaluate

$$\sum_{x \in L(2)} H[x] = \langle H, \sum_{x \in L(2)} x x^{\mathsf{T}} \rangle$$

and use the fact that A_1 and A_3 form spherical 2-designs and so

$$\sum_{x \in L(2)} x x^{\mathsf{T}} = 2h(A_1) I_8 \oplus 2h(A_3) I_{24} = 4 I_8 \oplus 8 I_{24}.$$

The matrix $H = 24 I_8 \oplus (-8) I_{24}$ has trace zero and gives

$$\sum_{x \in L(2)} H[x] = 24 \cdot 4 \cdot 8 - 8 \cdot 8 \cdot 24 = -4 \cdot 8 \cdot 24 = -768 \neq 0.$$

Now, by the eigenvalue bounds for $H[x]$ coming from the Rayleigh-Ritz principle, we find that

$$-8 = \lambda_{\min}(H) \leq \frac{H[x]}{\|x\|^2} \leq \lambda_{\max}(H) = 24.$$

This allows to organize summation over all lattice vectors of squared length at least 4 by shells

$$-8 \sum_{m \geq 2} a_m \cdot 2m \cdot e^{-\alpha(2m)} \leq \sum_{x \in L\setminus(\{0\}\cup L(2))} H[x] e^{-\alpha\|x\|^2} \leq 24 \sum_{m \geq 2} a_m \cdot 2m \cdot e^{-\alpha(2m)},$$

where $a_m = |L(2m)|$ ist the m-th coefficient of the theta series Θ_L of L.

Combining the above, we see that it suffices to show

$$24 \sum_{m \geq 2} a_m \cdot 2m \cdot e^{-\alpha(2m)} \leq 768 \cdot e^{-2\alpha}. \tag{6.33}$$

For this we write Θ_L in the form $\Theta_L = E_{16} + f$, where f is a cusp form of weight 16. Let

$$E_{16}(\tau) = \sum_{m=0}^{\infty} b_m q^m \quad \text{and} \quad f(\tau) = \sum_{m=1}^{\infty} c_m q^m,$$

in particular $b_m = -\frac{32}{B_{16}}\sigma_{15}(m) = 16320/3617\sigma_{15}(m)$ and so $c_1 = -16320/3617$. We use the estimate $\sigma_{k-1}(m) \leq \zeta(k-1)m^{k-1}$, where ζ is the Riemann zeta function, and get $b_m \leq 4.6m^{15}$. To bound c_m we use (6.4), the facts $\ell = 1$, $d(m) \leq 2\sqrt{m}$, and get $|c_m| \leq 1.2 \cdot 10^{10} m^8$. Together,

$$|a_m| \leq 4.6m^{15} + 1.2 \cdot 10^{10} m^8. \tag{6.34}$$

We evaluate for $\alpha = 14$, this gives

$$\sum_{m \geq 2} a_m \cdot 2m \cdot e^{-28m} \leq 9.2 \sum_{m=2}^{\infty} m^{16} \cdot e^{-28m} + 2.4 \cdot 10^{10} \sum_{m=2}^{\infty} m^9 \cdot e^{-28m}.$$

By Lemma 6.2.2 we have

$$\sum_{m=2}^{\infty} m^{16} e^{-28m} \leq 3.2 \cdot 10^{-20} + (28)^{-17}\Gamma(17, 56) \leq 3.3 \cdot 10^{-20},$$

and

$$\sum_{m=2}^{\infty} m^9 e^{-28m} \leq 2.5 \cdot 10^{-22} + (2\alpha)^{-10}\Gamma(10, 56) \leq 2.6 \cdot 10^{-22}.$$

Putting everything together for $\alpha = 14$ in (6.33) we find

$$\begin{aligned} 24 \sum_{m \geq 2} a_m \cdot 2m \cdot e^{-\alpha(2m)} &\leq 24\left(9.2 \cdot 3.3 \cdot 10^{-20} + 2.4 \cdot 10^{10} \cdot 2.6 \cdot 10^{-22}\right) \\ &\leq 24\left(3.1 \cdot 10^{-19} + 6.3 \cdot 10^{-12}\right) \\ &\leq 1.6 \cdot 10^{-10} \\ &\leq 768 \cdot e^{-28}. \end{aligned}$$

This shows that this lattice is not a critical point for the Gaussian potential function e^{-14r}.

Last, but not least, we show that all even unimodular lattices without roots in dimension 32 are local maxima for the Gaussian potential

function $e^{-\pi r}$. All the even unimodular lattices in dimension 32 without roots have the same theta series, for such a lattice $L \subseteq \mathbb{R}^{32}$ we have

$$\begin{aligned}\Theta_L(\tau) &= E_4^4(\tau) - 960E_4(\tau)\Delta(\tau)\\ &= 1 + 146880q^2 + 64757760q^3 + 4844836800q^4 + 137695887360q^5\\ &\quad + 2121555283200q^6 + 21421110804480q^7\\ &\quad + 158757684004800q^8 + \cdots\end{aligned}$$

All shells of L form spherical 4-designs, so L is critical for all Gaussian potential functions and we can compute the eigenvalue of the Hessian (6.7) using (6.12). For $\alpha = \pi$ we compute the first summands of the series and get

$$\frac{1}{n(n+2)} \sum_{m=0}^{8} a_m \pi(2m)(\pi(2m) - (n/2+1))e^{-\pi(2m)} < -0.00027.$$

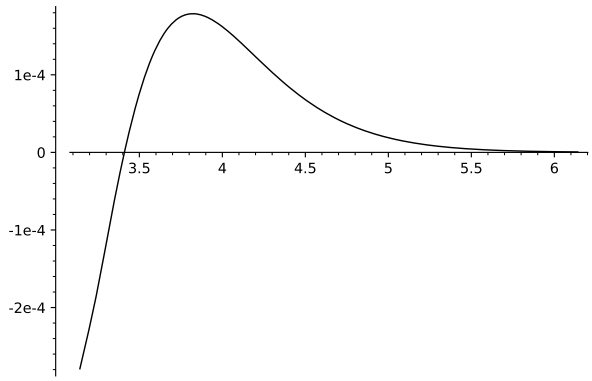


Figure 6.2: The eigenvalue of the Hessian for even unimodular lattices in dimension 32 without roots depending on the parameter α.

Now we argue that the tail of the series is so small that the entire series is strictly negative.

For this, again, we use the bound (6.34) for the coefficients a_m of Θ_L, and we estimate

$$\left| \sum_{m=9}^{\infty} a_m \pi(2m)(\pi(2m) - (n/2+1))e^{-\pi(2m)} \right| \leq \sum_{m=9}^{\infty} |a_m| (2\pi m)^2 e^{-2\pi m},$$

and

$$\sum_{m=9}^{\infty} |a_m|(2\pi m)^2 e^{-2\pi m} \leq 181.7 \sum_{m=9}^{\infty} m^{17} e^{-2\pi m} + 4.8 \cdot 10^{11} \sum_{m=9}^{\infty} m^{10} e^{-2\pi m}.$$

Again, by Lemma 6.2.2

$$\sum_{m=9}^{\infty} m^{17} e^{-2\pi m} \leq 4.7 \cdot 10^{-9} + (2\pi)^{-18} \Gamma(18, 18\pi) \leq 5.8 \cdot 10^{-9}.$$

Similarly,

$$\sum_{m=9}^{\infty} m^{10} e^{-2\pi m} \leq 9.7 \cdot 10^{-16} + (2\pi)^{-11} \Gamma(11, 18\pi) \leq 1.2 \cdot 10^{-15}.$$

Altogether:

$$\begin{aligned} &\left| \frac{1}{n(n+2)} \sum_{m=9}^{\infty} a_m \pi(2m)(\pi(2m) - (n/2+1)) e^{-\pi(2m)} \right| \\ &\leq \ 1088^{-1} \left(181.7 \cdot 5.8 \cdot 10^{-9} + 4.8 \cdot 10^{11} \cdot 1.2 \cdot 10^{-15} \right) \\ &\leq \ 5.4 \cdot 10^{-7}. \end{aligned}$$

Hence, we showed that for $\alpha = \pi$ the even unimodular lattices in dimension 32 without roots are local maxima for the Gaussian potential function. This answers a question of Regev and Stephens-Davidowitz [RSD17].

Acknowledgements

F.V. and M.C.Z. thank Noah Stephens-Davidowitz for a discussion at the Simons Institute for the Theory of Computing during the workshop "Lattices: Geometry, Algorithms and Hardness workshop" (February 18–21, 2020, organized by Daniele Micciancio, Daniel Dadush, Chris Peikert) which lead to the results of this paper.

We thank the anonymous referees for their helpful comments, suggestions, and corrections on the manuscript.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie agreement No 764759. F.V. is partially supported by the SFB/TRR

191 “Symplectic Structures in Geometry, Algebra and Dynamics”, F.V. and M.C.Z. are partially supported “Spectral bounds in extremal discrete geometry” (project number 414898050), both funded by the DFG. A.H. is partially supported by the DFG under the Priority Program CoSIP (project number SPP1798). A.T. is partially funded through HYPATIA.SCIENCE by the financial fund for the implementation of the statutory equality mandate and the Department of Mathematics and Computer Science, University of Cologne.

Chapter 7

A semidefinite program for least distortion embeddings of flat tori into Hilbert spaces

About this section

The following text has been previously published as:

> Arne Heimendahl, Moritz Lücke, Frank Vallentin, Marc Christian Zimmermann. "A semidefinite program for least distortion embeddings of flat tori into Hilbert spaces". arXiv: 2210.11952

Changes from the journal version are limited to typesetting and notation. These changes were performed to match the rest of this dissertation.

All authors contributed equally to this work. In particular, AH developed the proof of Theorem 7.4.2 and worked out the proofs of Section 7.7.

Abstract

We derive and analyze an infinite-dimensional semidefinite program which computes least distortion embeddings of flat tori $\mathbb{R}^n/L$, where L is an n-dimensional lattice, into Hilbert spaces.

This enables us to provide a constant factor improvement over the previously best lower bound on the minimal distortion of an embedding of an n-dimensional flat torus.

As further applications we prove that every n-dimensional flat torus has a finite dimensional least distortion embedding, that the standard embedding of the standard tours is optimal, and we determine least distortion embeddings of all 2-dimensional flat tori.

7.1 Introduction

Least distortion embeddings of flat tori into Hilbert spaces were first studied by Khot and Naor [KN05] in 2006. One motivation is that studying the Euclidean distortion of flat tori might have applications to the complexity of lattice problems, like the closest vector problem, and might also lead to more efficient algorithms for lattice problems through the use of least distortion embeddings. Another motivation comes from comparing the Riemannian setting to the bi-Lipschitz setting we are discussing here. On the one hand, by the Nash embedding theorem, flat tori can be embedded isometrically as Riemannian submanifolds into Euclidean space; we refer to [BJLT12] for spectacular visualizations of such an isometric embedding in the case of the two-dimensional square flat torus. On the other hand, Khot and Naor showed that flat tori can be highly non-Euclidean in the bi-Lipschitz setting.

7.1.1 Notation and review of the relevant literature

We review the relevant results of the literature which appeared since the pioneering work of Khot and Naor. At the same time we set the notation for this paper.

A flat torus is the metric space given by the quotient $\mathbb{R}^n/L$ with some n-dimensional lattice $L \subseteq \mathbb{R}^n$ and with metric

$$d_{\mathbb{R}^n/L}(x, y) = \min_{v \in L} |x - y - v|.$$

An n-dimensional lattice is a discrete subgroup of $(\mathbb{R}^n, +)$ consisting of all integral linear combinations of a basis of $\mathbb{R}^n$. Furthermore, $|\cdot|$ denotes the standard norm of $\mathbb{R}^n$ given by $|x| = \sqrt{x^\mathsf{T} x}$.

A Euclidean embedding of $\mathbb{R}^n/L$ is an injective function $\varphi\colon \mathbb{R}^n/L \to H$ mapping the flat torus $\mathbb{R}^n/L$ into some (complex) Hilbert space H.

The *distortion* of φ is

$$\operatorname{dist}(\varphi) = \sup_{\substack{x,y\in\mathbb{R}^n/L\\ x\neq y}} \frac{\|\varphi(x)-\varphi(y)\|}{d_{\mathbb{R}^n/L}(x,y)} \cdot \sup_{\substack{x,y\in\mathbb{R}^n/L\\ x\neq y}} \frac{d_{\mathbb{R}^n/L}(x,y)}{\|\varphi(x)-\varphi(y)\|},$$

where $\|\cdot\|$ is the norm of the Hilbert space H. Here the first supremum is called the *expansion* of φ and the second supremum is the *contraction* of φ. When we minimize the distortion of φ over all possible embeddings of $\mathbb{R}^n/L$ into Hilbert spaces we speak of the *least (Euclidean) distortion* of the flat torus; it is denoted by

$$c_2(\mathbb{R}^n/L) = \inf\{\operatorname{dist}(\varphi) \ : \ \varphi\colon \mathbb{R}^n/L \to H \text{ for some Hilbert space } H,\ \varphi \text{ injective}\}.$$

Similarly one can define $c_1(\mathbb{R}^n/L)$ by replacing the Hilbert space by some L_1 space.

Khot and Naor showed (see [KN05, Corollary 4]) that flat tori can be highly non-Euclidean in the sense that there is a family of flat tori $\mathbb{R}^n/L_n$ with

$$c_2(\mathbb{R}^n/L_n) = \Omega(\sqrt{n}). \tag{7.1}$$

On the other hand, they noticed (see [KN05, Remark 5]) that the standard embedding of the standard flat torus $\mathbb{R}^n/\mathbb{Z}^n$ embeds into $\mathbb{R}^{2n}$ with distortion $O(1)$[1]. The standard embedding is given by

$$\varphi(x_1,\dots,x_n) = (\cos 2\pi x_1, \sin 2\pi x_1, \dots, \cos 2\pi x_n, \sin 2\pi x_n). \tag{7.2}$$

In fact, Khot and Naor are mainly concerned with bounding $c_1(\mathbb{R}^n/L)$, which immediately provides bounds for $c_2(\mathbb{R}^n/L)$ because $c_1(\mathbb{R}^n/L) \le c_2(\mathbb{R}^n/L)$. To state their main result, leading to (7.1), we make use of the Voronoi cell of L, which is an n-dimensional polytope defined as

$$V(L) = \{x \in \mathbb{R}^n : |x| \le |x-v| \text{ for all } v \in L\}.$$

The Voronoi cell is a fundamental domain of $\mathbb{R}^n/L$ under the action of L. We denote the volume of $V(L)$ by $\operatorname{vol} L$. Clearly, $|x| = d_{\mathbb{R}^n/L}(x,0)$ for all $x \in V(L)$. The covering radius of L is $\mu(L) = \max\{|x| : x \in V(L)\}$, which is the circumradius of $V(L)$. The length of a shortest

[1] In fact, we have $\operatorname{dist}(\varphi) = \pi/2$ and φ is an optimal embedding, see Theorem 7.6.1.

vector of L is $\lambda(L) = \min\{|v| : v \in L \setminus \{0\}\}$ that is two times the inradius of $V(L)$. Now the main result (see [KN05, Theorem 5]) is

$$c_1(\mathbb{R}^n/L) = \Omega\left(\frac{\lambda(L^*)\sqrt{n}}{\mu(L^*)}\right). \tag{7.3}$$

Here, as usual, $L^* = \{u \in \mathbb{R}^n : u^\mathsf{T} v \in \mathbb{Z} \text{ for all } v \in L\}$ denotes the dual lattice of L. They also give an alternative proof of their main result for $c_2(\mathbb{R}^n/L)$ (see [KN05, Lemma 11]). The main result leads to the lower bound (7.1) when plugging in duals of lattices which simultaneously provide dense packings and economical coverings. Such a family of lattices exist by a theorem of Butler [But72].

Using Korkine-Zolotarev reduction Khot and Naor determine an embedding of $\mathbb{R}^n/L$ into $\mathbb{R}^{2n}$ with distortion $O(n^{3n/2})$ (see [KN05, Theorem 6]).

Haviv and Regev [HR13, Theorem 1.3] found an improved embedding that yields $c_2(\mathbb{R}^n/L) = O(n\sqrt{\log n})$. They also improved on (7.3) and showed in [HR13, Theorem 1.5] that for any n-dimensional lattice L we have

$$c_2(\mathbb{R}^n/L) \geq \frac{\lambda(L^*)\mu(L)}{4\sqrt{n}}, \tag{7.4}$$

which improves on (7.3) because $\mu(L)\mu(L^*) \geq \Omega(n)$ holds for every n-dimensional lattice. This follows from a simple volume argument giving $\mu(L) = \Omega(\sqrt{n}(\operatorname{vol} L)^{1/n})$ and $\operatorname{vol} L^* = (\operatorname{vol} L)^{-1}$.

Recently, Agarwal, Regev, Tang [ART20] constructed excellent embeddings of flat tori having low distortion and showed that the lower bound (7.1) is nearly tight: For every lattice $L \subseteq \mathbb{R}^n$ there exists an embedding of $\mathbb{R}^n/L$ into Hilbert space with distortion $O(\sqrt{n \log n})$.

7.1.2 Aim and method

In this paper we want to add a semidefinite optimization perspective to this story.

For finite metric spaces it is known that one can compute least distortion Euclidean embeddings via a semidefinite program (SDP), which is linear optimization over the cone of positive semidefinite matrices. We want to extend this result from finite metric spaces to

flat tori. This will yield, via semidefinite programming duality, an algorithmic method for proving nonembeddability results. In particular, this leads to a new, simple proof of (7.4). In fact we even get a constant factor improvement that is tight in the case of the standard torus.

First we recall the semidefinite program for finding least Euclidean distortion embeddings of finite metric spaces. Suppose we consider a finite metric space X with distance function d. Then, as first observed by Linial, London, Rabinovich [LLR95], we can find a least distortion embedding of (X, d) into a Hilbert space algorithmically by solving the following semidefinite program

$$\begin{aligned}\min\{C \ : C \in \mathbb{R}_+, Q \in \mathcal{S}_+^X,\\ d(x,y)^2 \le Q_{xx} - 2Q_{xy} + Q_{yy} \le Cd(x,y)^2 \text{ for all } x, y \in X\},\end{aligned} \tag{7.5}$$

where $\mathcal{S}_+^X$ denotes the convex cone of positive semidefinite matrices whose rows and columns are indexed by the elements of X. The optimal solution C of this semidefinite program equals $c_2(X, d)^2$ and if Q attains the optimal solution, then we can determine a least distortion embedding $\varphi : X \to \mathbb{R}^X$ with the property $\varphi(x) \cdot \varphi(y) = Q_{xy}$ by considering a Cholesky decomposition of Q.

This shows how to compute (in fact in polynomial time) an optimal Euclidean embedding of a finite metric space. Another benefit of this formulation is that we can apply duality theory of semidefinite programs. Then the dual maximization problem will play a key role to determine lower bounds for $c_2(X, d)$. By using strong duality we arrive at the following result: The least distortion of a finite metric space (X, d), with $X = \{x_1, \ldots, x_n\}$, into Euclidean space is given by

$$c_2(X,d)^2 = \max\left\{\frac{\sum_{i,j=1:Y_{ij}>0}^n Y_{ij} d(x_i,x_j)^2}{-\sum_{i,j=1:Y_{ij}<0}^n Y_{ij} d(x_i,x_j)^2} : Y \in \mathcal{S}_+^n, Y\mathbf{e} = 0\right\}. \tag{7.6}$$

The condition $Y\mathbf{e} = 0$ says that the all-ones vector $\mathbf{e}$ lies in the kernel of Y. A proof of this result is detailed in Matoušek [Mat02] or in Laurent, Vallentin [LV].

This lower bound has been extensively used to determine the least distortion Euclidean embeddings of the shortest path metric of several graph classes. Linial, Magen [LM00] computed least distortion embeddings of products of cycles and of expander graphs. Least distortion Euclidean embeddings of strongly regular graphs and of more general distance regular graphs were first considered by Vallentin [Val08]. This was further extended by Kobayashi, Kondo [KK15], Cioabă, Gupta, Ihringer, Kurihara [CGIK21]. Linial, Magen, Naor [LMN02] considered graphs of high girth using this approach.

To apply the bound (7.6) one has to construct a matrix Y, which sometimes appears to come out of the blue. By complementary slackness, which is the same as analyzing the case of equality in the proof of weak duality, we get hints where to search for an appropriate matrix Y: If Y is an optimal solution of the maximization problem (7.6), then $Y_{ij} > 0$ only for the *most contracted pairs*. These are pairs (x_i, x_j) for which $\frac{d(x_i,x_j)}{\|f(x_i)-f(x_j)\|}$ is maximized. Similarly, then $Y_{ij} < 0$ only for the *most expanded pairs*, maximizing $\frac{\|f(x_i)-f(x_j)\|}{d(x_i,x_j)}$.

Linial, Magen [LM00] realized that for graphs most expanded pairs are simply adjacent vertices. However, most contracted pairs are more mysterious and there is no characterization known. The first intuition that the largest contraction occurs at pairs at maximum distance is wrong in general.

7.1.3 Contribution and structure of the paper

In Section 7.2 of this paper we derive a new infinite-dimensional semidefinite program for determining a least distortion embedding of flat tori into Hilbert spaces which is analogous to (7.5). It is given in Theorem 7.2.1 where we additionally apply symmetry reduction techniques in the spirit of [BGSV12] to reduce the original infinite-dimensional SDP into an infinite-dimensional linear program that involves Fourier analysis. Then we realize that in a Euclidean embedding of a flat torus there are no most expanded pairs: The expansion is only attained in the limit by pairs whose distance tends to zero. This is in perfect analogy to the graph case where the most expanded pairs are

also attained at minimal distance. This insight has the advantage that in the infinite dimensional linear program some of the infinitely many constraints can be replaced by only one finite-dimensional semidefinite constraint. This is the content of Theorem 7.2.4. Its dual program is derived in Theorem 7.2.5 which is analogous to (7.6).

In Section 7.3 we further investigate the properties of the optimization problems given in Theorem 7.2.4 and Theorem 7.2.5. These properties will be used in the next sections.

In the last sections we apply our new methodology. In Section 7.4 we prove that an n-dimensional flat torus always admits a finite dimensional least distortion embedding, a space of (complex) dimension $2^n - 1$ suffices. Section 7.5 contains a new and simple proof of our constant factor improvement of the lower bound given in (7.4). In Section 7.6 we show that the standard embedding (7.2) of the standard torus is indeed optimal and has distortion $\pi/2$. In Section 7.7 we determine least distortion embeddings of all two-dimensional flat tori. A few open questions are discussed in Section 7.8.

7.2 An infinite-dimensional SDP

Starting from (7.5) we want to derive a similar, but now infinite-dimensional, semidefinite program which can be used to determine $c_2(\mathbb{R}^n/L)$.

7.2.1 Primal program

The first step is to apply a classical theorem of Moore [Moo16] which enables us to optimize over all embeddings $\varphi : \mathbb{R}^n/L \to H$ into some Hilbert space H. In our situation Moore's theorem says that there exists a (complex) Hilbert space H and a map $\varphi : \mathbb{R}^n/L \to H$ if and only if there is a positive definite kernel[2]

$$Q : \mathbb{R}^n/L \times \mathbb{R}^n/L \to \mathbb{C} \text{ such that } Q(x,y) = (\varphi(x), \varphi(y)) \text{ for all } x, y \in \mathbb{R}^n/L,$$

[2] A kernel Q is called positive definite if and only if for all $N \in \mathbb{N}$ and for all $x_1, \ldots, x_N \in \mathbb{R}^n/L$ the matrix $(Q(x_i, x_j))_{1 \le i,j \le \mathbb{N}} \in \mathbb{C}^{N \times N}$ is Hermitian and positive semidefinite. This naming convention is unfortunate but for historical reasons unavoidable.

where $(\cdot,\cdot)$ denotes the inner product of H. Therefore we get

$$\begin{aligned} c_2(\mathbb{R}^n/L)^2 = \inf\{C \ : C \in \mathbb{R}_+, Q \text{ positive definite,} \\ d_{\mathbb{R}^n/L}(x,y)^2 \ &\leq \ Q(x,x) - 2\,\mathrm{Re}(Q(x,y)) + Q(y,y) \\ &\leq \ C d_{\mathbb{R}^n/L}(x,y)^2 \ \text{ for all } x,y \in \mathbb{R}^n/L\}. \end{aligned}$$

Here we scaled the embedding φ which is defined through Q so that the contraction of φ equals 1. The real part $\mathrm{Re}(Q)$ of a positive definite kernel is positive definite again and we can restrict to real-valued positive definite kernels for determining $c_2(\mathbb{R}^n/L)$.

For the second step we apply a standard group averaging argument. If Q is a feasible solution for the minimization problem above, so is its group average

$$\overline{Q}(x,y) = \frac{1}{\mathrm{vol}(\mathbb{R}^n/L)} \int_{\mathbb{R}^n/L} Q(x-z, y-z)\, dz.$$

By this averaging the kernel $\overline{Q}$ becomes continuous and only depends on the difference $x-y$. Thus, instead of minimizing over positive definite kernels Q it suffices to minimize over continuous, real functions $f\colon \mathbb{R}^n/L \to \mathbb{R}$ which are of positive type, i.e. the kernel $(x,y) \mapsto f(x-y)$ is positive definite; see also the proof of Theorem 3.1 in the paper [AMM85] by Aharoni, Maurey, Mityagin.

For the convenience of the reader we provide the argument why the positive type function $f(x) = \overline{Q}(x,0)$ is continuous: For every x, y the matrix

$$\begin{pmatrix} f(0) & f(x) & f(x+y) \\ f(x) & f(0) & f(y) \\ f(x+y) & f(y) & f(0) \end{pmatrix}$$

is positive semidefinite and it is congruent (simultaneously subtract the second row/column of the third row/column) to the positive semidefinite matrix

$$\begin{pmatrix} f(0) & f(x) & f(x+y)-f(x) \\ f(x) & f(0) & f(y)-f(0) \\ f(x+y)-f(x) & f(y)-f(0) & 2f(0)-2f(y) \end{pmatrix}.$$

Taking the minor of the first and third row/column gives

$$2f(0)(f(0)-f(y)) \geq (f(x+y)-f(x))^2.$$

This inequality implies that f is continuous at every x if and only if f is continuous at 0. Then f is continuous at 0 because it satisfies the constraint

$$d_{\mathbb{R}^n/L}(0,y)^2 \le \overline{Q}(0,0)-2\overline{Q}(0,y)+\overline{Q}(y,y) = 2(f(0)-f(y)) \le C d_{\mathbb{R}^n/L}(0,y)^2$$

for every y.

Note that also $(x,y) \mapsto d_{\mathbb{R}^n/L}(x,y)^2$ only depends on the difference $x-y$. So we can replace (x,y) by $(x-y,0)$ and we can move $x-y$ by a lattice vector translation into the Voronoi cell $V(L)$. Hence,

$$\begin{aligned} c_2(\mathbb{R}^n/L)^2 = \inf\{C \ : C \in \mathbb{R}_+, f\colon \mathbb{R}^n/L \to \mathbb{R} \text{ continuous and of positive type,} \\ |x|^2 \le 2(f(0)-f(x)) \le C|x|^2 \ \text{ for all } x \in V(L)\}. \end{aligned}$$

In the third step we parametrize continuous positive type functions by the Fourier coefficients using Bochner's theorem, cf. Folland [Fol95, (4.18)], which says that a continuous function $f : \mathbb{R}^n/L \to \mathbb{C}$ is of positive type if and only if all its Fourier coefficients

$$\widehat{f}(u) = \int_{\mathbb{R}^n/L} f(x) e^{-2\pi i u^{\mathsf{T}} x}\, dx,$$

with $u \in L^*$ are nonnegative and $\widehat{f}$ lies in

$$\ell^1(L^*) = \left\{ z\colon L^* \to \mathbb{C} : \sum_{u \in L^*} |z(u)| < \infty \right\}.$$

Then if f is real, continuous and of positive type we have the representation

$$f(x) = \sum_{u \in L^*} \widehat{f}(u) e^{2\pi i u^{\mathsf{T}} x},$$

where the convergence is absolute and uniform, with $\widehat{f} \in \ell^1(L^*)$, $\widehat{f}(u) \ge 0$ and $\widehat{f}(u) = \widehat{f}(-u)$ for all $u \in L^*$. Thus,

$$f(x) = \sum_{u \in L^*} \widehat{f}(u) \cos(2\pi u^{\mathsf{T}} x).$$

Writing f in this form, one can express $c_2(\mathbb{R}^n/L)^2$ as an infinite-dimensional linear program:

Theorem 7.2.1. *The least distortion Euclidean embedding of a flat torus $\mathbb{R}^n/L$ is given by*

$$c_2(\mathbb{R}^n/L)^2 = \inf\big\{C \ : C \in \mathbb{R}_+, z \in \ell^1(L^*), z(u) = z(-u) \geq 0 \text{ for all } u \in L^*, \\ |x|^2 \leq 2\sum_{u\in L^*} z(u)(1-\cos(2\pi u^{\mathsf{T}}x)) \leq C|x|^2 \\ \text{for all } x \in V(L)\big\}. \tag{7.7}$$

A feasible solution of the above minimization problem (C, z) determines a Euclidean embedding φ of $\mathbb{R}^n/L$ with distortion $dist(\varphi) \leq \sqrt{C}$ by

$$\varphi : \mathbb{R}^n/L \to \ell^2(L^*), \quad x \mapsto \left(\sqrt{z(u)}e^{2\pi i u^{\mathsf{T}}x}\right)_{u\in L^*}, \tag{7.8}$$

with complex Hilbert space

$$\ell^2(L^*) = \left\{z : L^* \to \mathbb{C} : \left(\sum_{u\in L^*} |z(u)|^2\right)^{1/2} < \infty\right\}.$$

Remark 7.2.2. *The* inf *in* (7.7) *is in fact a* min *because the set of bounded continuous functions of positive type is weak* compact due to the Banach-Alaoglu theorem; see for example Folland [Fol95, Chapter 3.3].*

It is worth to mention that the embedding φ of Theorem 7.2.1 embeds the flat torus $\mathbb{R}^n/L$ into a direct product of circles

$$\prod_{u\in L^*} \sqrt{z(u)}S^1 \quad \text{with} \quad \|\varphi(x)\|^2 = \sum_{u\in L^*} z(u) \text{ for all } x \in L.$$

The support of z contains a basis of L^* since the embedding is injective. Using the fact $z(u) = z(-u)$ we could also use the real embedding φ' with

$$[\varphi'(x)]_u = \sqrt{z(u)}(\cos 2\pi u^{\mathsf{T}}x, \sin 2\pi u^{\mathsf{T}}x),$$

where u runs through $L^*/\{\pm 1\}$ and which has the same distortion as φ.

On the other hand, the constraint $z(u) = z(-u)$ is clearly redundant in the minimization problem of Theorem 7.2.1.

Now we want to simplify the infinitely many inequalities

$$2 \sum_{u \in L^*} z(u)(1 - \cos(2\pi u^\mathsf{T} x)) \leq C|x|^2 \text{ for all } x \in V(L), \tag{7.9}$$

which occur in (7.7), by only *one* finite-dimensional semidefinite condition. For this we observe that in any embedding there are no most expanded pairs: the corresponding supremum $\sup\left\{\frac{\|\varphi(x)-\varphi(y)\|}{d_{\mathbb{R}^n/L}(x,y)} : x, y \in \mathbb{R}^n/L, x \neq y\right\}$ is only attained by a limit of pairs whose distance tends to 0.

Lemma 7.2.3. *Let $L \subseteq \mathbb{R}^n$ be an n-dimensional lattice. Let (C, z) be as in* (7.7)*. Inequality* (7.9) *is satisfied if and only if*

$$4\pi^2 \sum_{u \in L^*} z(u)(u^\mathsf{T} x)^2 \leq C|x|^2 \text{ for all } x \in \mathbb{R}^n. \tag{7.10}$$

Note that (9) holds for all $x \in \mathbb{R}^n$.

Proof. By the cosine double angle formula $1 - \cos(\alpha) = 2\sin(\alpha/2)^2$ and by the inequality $|\sin(\alpha)| \leq |\alpha|$ we have

$$2 \sum_{u \in L^*} z(u)(1 - \cos(2\pi u^\mathsf{T} x)) \leq 4\pi^2 \sum_{u \in L^*} z(u)(u^\mathsf{T} x)^2.$$

Thus, (7.10) implies (7.9).

Conversely, assume that (7.10) is not satisfied. There exists $x^* \in \mathbb{R}^n$ with

$$4\pi^2 \sum_{u \in L^*} z(u)(u^\mathsf{T} x^*)^2 > C|x^*|^2.$$

For $r \geq 0$ define the function

$$f(r) = 2 \sum_{u \in L^*} z(u)(1 - \cos(2\pi u^\mathsf{T}(rx^*))) - C|rx^*|^2$$

and consider its Taylor expansion

$$f(r) = \left(4\pi^2 \sum_{u \in L^*} z(u)(u^\mathsf{T} x^*)^2 - C|x^*|^2\right) r^2 + \text{h.o.t. (in } r)$$

Writing f this way and using the assumption, $f(r)$ is positive for sufficiently small r. Thus, (7.9) is not satisfied. □

Inequality (7.10) can also be rewritten as an inequality of the largest eigenvalue $\lambda_{\max}$ of a corresponding matrix

$$\lambda_{\max}\left(4\pi^2 \sum_{u\in L^*} z(u)uu^{\mathsf{T}}\right) \leq C$$

or equivalently as a semidefinite condition

$$CI - 4\pi^2 \sum_{u\in L^*} z(u)uu^{\mathsf{T}} \in \mathcal{S}^n_+,$$

where I denotes the identity matrix. With this lemma we arrive at the following simplification of (7.7).

Theorem 7.2.4. *The least distortion Euclidean embedding of a flat torus $\mathbb{R}^n/L$ is given by*

$$\begin{aligned} c_2(\mathbb{R}^n/L)^2 = \inf\big\{C \ : C \in \mathbb{R}_+, z \in \ell^1(L^*), z(u) = z(-u) \geq 0 \text{ for all } u \in L^*,\\ |x|^2 \leq 2\sum_{u\in L^*} z(u)(1-\cos(2\pi u^{\mathsf{T}}x)) \text{ for all } x \in V(L),\\ CI - 4\pi^2 \sum_{u\in L^*} z(u)uu^{\mathsf{T}} \in \mathcal{S}^n_+\big\}. \end{aligned} \tag{7.11}$$

7.2.2 Dual program

We derive the dual of (7.11) to systematically find lower bounds for $c_2(\mathbb{R}^n/L)$.

Theorem 7.2.5. *Suppose that (C, z) is feasible for (7.11), then*

$$\begin{aligned} C \geq c_2(\mathbb{R}^n/L)^2 \geq \sup\big\{2\pi^2 \int_{V(L)} |x|^2 \, d\nu(x) \ :\\ \nu \in \mathcal{M}_+(V(L)), Y \in \mathcal{S}^n_+, \mathrm{Tr}(Y) = 1,\\ \int_{V(L)} (1-\cos(2\pi u^{\mathsf{T}}x))\, d\nu(x) \leq u^{\mathsf{T}}Yu\\ \text{for all } u \in L^*\big\}, \end{aligned} \tag{7.12}$$

where $\mathcal{M}_+(V(L))$ is the cone of Borel measures supported on $V(L)$. In (7.12) equality holds for a feasible (ν, Y) if and only if

$$\left(CI - 4\pi^2 \sum_{u\in L^*} z(u)uu^{\mathsf{T}}\right) Y = 0,$$

and the measure ν is only supported on vectors $x \in V(L)$ for which equality

$$|x|^2 = 2 \sum_{u \in L^*} z(u)(1 - \cos(2\pi u^\mathsf{T} x))$$

holds, and for all vectors $u \in L^*$ with $z(u) \neq 0$ we have

$$\int_{V(L)} (1 - \cos(2\pi u^\mathsf{T} x))\, d\nu(x) = u^\mathsf{T} Y u.$$

Proof. For two symmetric matrices A, B we define $\langle A, B \rangle = \mathrm{Tr}(AB)$. Using the feasibility of (C, z) and (ν, Y) we get

$$\begin{aligned}
&C - 2\pi^2 \int_{V(L)} |x|^2\, d\nu(x) \\
&\geq \left\langle 4\pi^2 \sum_{u \in L^*} z(u) u u^\mathsf{T}, Y \right\rangle - 4\pi^2 \int_{V(L)} \sum_{u \in L^*} z(u)(1 - \cos(2\pi u^\mathsf{T} x))\, d\nu(x) \\
&= 4\pi^2 \sum_{u \in L^*} z(u) \left(\langle u u^\mathsf{T}, Y \rangle - \int_{V(L)} (1 - \cos(2\pi u^\mathsf{T} x))\, d\nu(x) \right) \\
&\geq 0.
\end{aligned}$$

When analyzing the case of equality we find the three conditions of the theorem. □

Remark 7.2.6. *As a side note we would like to mention that in* (7.12) *we even have equality $c_2(\mathbb{R}^n/L)^2 = \sup$. This follows again by the weak* compactness of the set of bounded, continuous functions of positive type together with the Hahn-Banach (strict) separation theorem.*

7.3 Properties and observations

We collect some results that are consequences of the primal and dual formulation of the preceding section, including some auxiliary results used in later sections.

7.3.1 Subquadratic inequality

First, we show that the functions of the form

$$f(x) = 2 \sum_{u \in L^*} z(u)(1 - \cos(2\pi u^\mathsf{T} x)) \text{ with } z(u) \geq 0 \tag{7.13}$$

are subquadratic, this auxiliary result is going to be used a number of times. Note that we have

$$f(x-y) = \|\varphi(x) - \varphi(y)\|^2$$

for the embedding φ in (7.8). Suppose for a moment that φ was an isometry, then f would satisfy the parallelogram law

$$f(x-y) + f(x+y) = 2f(x) + 2f(y)$$

and it would be a homogeneous quadratic form

$$f(\lambda x) = \lambda^2 f(x).$$

However, φ cannot be a Hilbert space isometry, but the next lemma shows that we have at least two inequalities.

Lemma 7.3.1. *The function f defined in* (7.13) *is subquadratic, i.e. it satisfies*

$$f(x+y) + f(x-y) \leq 2f(x) + 2f(y) \quad \textit{for all } x, y \in \mathbb{R}^n. \tag{7.14}$$

Furthermore,

$$f(\lambda x) \leq \lambda^2 f(x) \quad \textit{for all } \lambda \in \mathbb{N}, x \in \mathbb{R}. \tag{7.15}$$

If f defines an embedding, we have equality in (7.14) *and* (7.15) *if and only if x or y lie in L.*

A proof for (7.15) can also be found in [KTP06]; we provide it here for the convenience of the reader.

Proof. To show that f is subquadratic it suffices to prove the inequality

$$1 - \cos(\alpha+\beta) + 1 - \cos(\alpha-\beta) \leq 2(1-\cos\alpha) + 2(1-\cos\beta)$$

for all $\alpha, \beta \in \mathbb{R}$. This is elementary by the cosine addition formula:

$$\begin{aligned} 1 - \cos(\alpha+\beta) + 1 - \cos(\alpha-\beta) &= 2 - 2\cos\alpha\cos\beta \\ &= 2\cos\beta(1-\cos\alpha) + 2(1-\cos\beta) \\ &\leq 2(1-\cos\alpha) + 2(1-\cos\beta), \end{aligned}$$

where equality holds if and only if α or β is an integral multiple of 2π.

Now consider f as in (7.13) and assume f defines an embedding. Then the claim about equality comes from the fact that $\alpha = 2\pi u^\mathsf{T} x$ or $\beta = 2\pi u^\mathsf{T} y$ is an integral multiple of 2π for all $u \in \operatorname{supp}(z)$ if and only if $x \in L$ or $y \in L$, since $\operatorname{supp}(z)$ contains a basis of L^*.

For even λ we directly use (7.14)

$$\begin{aligned} f(\lambda x) = &\ f\left(\frac{\lambda}{2}x + \frac{\lambda}{2}x\right) + f\left(\frac{\lambda}{2}x - \frac{\lambda}{2}x\right) \\ &\leq 4\left(\frac{\lambda}{2}\right)^2 f(x) = \lambda^2 f(x) \end{aligned}$$

since $f(0) = 0$. For odd $\lambda \geq 3$ we use (7.14) and proceed by induction

$$\begin{aligned} f(\lambda x) + f(x) = &\ f\left(\left(\frac{\lambda-1}{2}+1\right)x + \frac{\lambda-1}{2}x\right) + f\left(\left(\frac{\lambda-1}{2}+1\right)x - \frac{\lambda-1}{2}x\right) \\ \leq &\ 2f\left(\left(\frac{\lambda-1}{2}+1\right)x\right) + 2f\left(\frac{\lambda-1}{2}x\right) \\ \leq &\ 2\left(\left(\frac{\lambda-1}{2}+1\right)^2 + \left(\frac{\lambda-1}{2}\right)^2\right) f(x) \\ = &\ \lambda^2 f(x) + f(x). \end{aligned}$$

□

7.3.2 Dual feasibility

In general the dual program (7.12) has infinitely many conditions of the form

$$\int_{V(L)} (1 - \cos(2\pi v^\mathsf{T} x))\, d\nu(x) \leq \operatorname{Tr}(vv^\mathsf{T} Y), \qquad v \in L^*. \tag{7.16}$$

We will now show that sometimes already finitely many constraints are sufficient to imply all conditions (7.16). The first observation is the following:

Lemma 7.3.2. *Let $q_a(x) = 1 - \cos(2\pi a^\mathsf{T} x)$. The (in-)equalities*

$$\int_{V(L)} q_a(x)\, d\nu(x) \leq \operatorname{Tr}(aa^\mathsf{T} Y), \quad \int_{V(L)} q_b(x)\, d\nu(x) \leq \operatorname{Tr}(bb^\mathsf{T} Y), \tag{7.17}$$

$$\int_{V(L)} q_{a-b}(x)\, d\nu(x) = \operatorname{Tr}((a-b)(a-b)^\mathsf{T} Y) \tag{7.18}$$

imply

$$\int_{V(L)} q_{a+b}(x)\,d\nu(x) \le \operatorname{Tr}((a+b)(a+b)^{\mathsf{T}}Y).$$

The corresponding result also holds when q_{a-b} and q_{a+b} are interchanged.

Proof. As shown in the proof of Lemma 7.3.1, the function q_a is subquadratic and therefore

$$\begin{aligned}\int_{V(L)} q_{a+b}(x)\,d\nu(x) + \int_{V(L)} q_{a-b}(x)\,d\nu(x) &\le 2\int_{V(L)} q_a(x) + q_b(x)\,d\nu(x)\\ &\le 2\operatorname{Tr}(aa^{\mathsf{T}}Y) + 2\operatorname{Tr}(bb^{\mathsf{T}}Y),\end{aligned}$$

which by (7.18) is equivalent to

$$\begin{aligned}\int_{V(L)} q_{a+b}(x)\,d\nu(x) &\le 2\operatorname{Tr}(aa^{\mathsf{T}}Y) + 2\operatorname{Tr}(bb^{\mathsf{T}}Y) - \operatorname{Tr}((a-b)(a-b)^{\mathsf{T}}Y)\\ &= \operatorname{Tr}((a+b)(a+b)^{\mathsf{T}}Y).\end{aligned}$$

□

The above lemma can be used to replace the infinitely many constraints (7.16) by finitely many using the shortest vectors in cosets of the form $v + 2L^*$ for $v \in L^*$.

The proof of the lemma relies on a characterization of *Voronoi vectors*. These are lattice vectors $v \in L \setminus \{0\}$ such that the set $F_v := V(L) \cap \{x : v^{\mathsf{T}}x \le \frac{1}{2}v^{\mathsf{T}}v\}$ defines a non-empty face of $V(L)$. Moreover, $v \in L$ is called *Voronoi relevant* if F_v is a facet of $V(L)$, i.e. an $(n-1)$-dimensional face of $V(L)$.

An element $v \in L \setminus \{0\}$ is a Voronoi vector of $V(L)$ if and only if $\pm v$ are shortest vectors in the coset $v + 2L$ and $\pm v$ are Voronoi relevant if and only if they are the *only* shortest vectors in $v + 2L$. For a proof see [CS88, Chapter 21, Theorem 10] and [CS92, Theorem 2].

Lemma 7.3.3. *If* (7.16) *is tight for at least one shortest vector in each coset of the form $v + 2L^*$, $v \in L^*$, then* (7.16) *holds for all $v \in L^*$.*

Proof. Assume that (7.16) is tight for at least one shortest vector in each coset $v + 2L^*$. We will first prove by induction that (7.16) also holds for all Voronoi vectors.

The statement holds by assumption for all Voronoi relevant vectors v because in this case the normal vectors v and $-v$ are the only shortest vectors in $v + 2L^*$ due to the above characterization.

So assume that the statement holds for all vectors $u \in L^*$ such that F_u is a face of dimension smaller than $n - k$. Now consider $v \in L^*$ such that v is the normal vector of some $(n-(k+1))$-dimensional face of $V(L^*)$ with $k \geq 2$. Further, assume that $u \in v + 2L^*$ with $u \neq v$ is such that (7.16) is tight for u (otherwise there is nothing to show for v).

Now $\frac{1}{2}(u \pm v) \in L^*$ and the inequality $v^\mathsf{T} x \leq \frac{1}{2} v^\mathsf{T} v$ for all $x \in V(L^*)$ is implied by

$$\begin{aligned} \frac{1}{2}(u+v)^\mathsf{T} x + \frac{1}{2}(u-v)^\mathsf{T} x \;&\leq\; \frac{1}{4}(u+v)^\mathsf{T}(u+v) + \frac{1}{4}(u-v)^\mathsf{T}(u-v) \\ &\leq\; \frac{1}{4}u^\mathsf{T} u + \frac{1}{4}v^\mathsf{T} v \;\leq\; \frac{1}{2}v^\mathsf{T} v. \end{aligned}$$

Due to $u \neq v$ and the above inequality, the sets $F_{\frac{1}{2}(u+v)}, F_{\frac{1}{2}(u-v)}$ define non-empty faces of $V(L^*)$ of dimension strictly larger than $n-(k+1)$. Hence, (7.16) holds for $\frac{1}{2}(u+v), \frac{1}{2}(u-v)$ by the induction hypothesis. Applying Lemma 7.3.2 (with $a = \frac{1}{2}(u+v)$, $b = \frac{1}{2}(u-v)$, $a+b = u$) shows that (7.16) also holds for v.

Now assume that v is a not a Voronoi vector. Then there exists a a shortest vector $u \in v + 2L^*$ for which (7.16) is tight and $|u| < |v|$.

Then, again $\frac{1}{2}(u \pm v) \in L^*$ and as

$$|\frac{1}{2}(u \pm v)| \leq \frac{1}{2}(|u| + |v|) < |v|,$$

we can argue by an analogous inductive argument (based on the norm) as before that (7.16) holds for $\frac{1}{2}(u \pm v)$. Finally, we can use Lemma 7.3.2 to infer that (7.16) is valid for v as well. □

7.4 Least Euclidean distortion embeddings always have finite dimension

The goal of this section is to prove that for every n-dimensional lattice, there always exists a least distortion embedding of $\mathbb{R}^n/L$ that is finite-dimensional. In the sense of Theorem 7.2.1, this means that there is

always an optimal solution (C, z) for (7.11) such that the support of z is finite.

Additionally, our arguments will reveal that the constructed optimal solution with finite support has only support on at most one vector per coset $v + 2L^*$ of $L^*/2L^*$ and that $\mathrm{supp}(z)$ only contains primitive lattice vectors. An element $v \in L$ is called *primitive* for L if $\alpha v \in L$ with $\alpha \in \mathbb{Z}$ implies $\alpha = \pm 1$.

The first step towards proving that there is always a finite-dimensional least Euclidean distortion embedding is the following observation.

Lemma 7.4.1. *Assume that (C, z) is a solution for (7.11).*

1. *If there are $u, v \in \mathrm{supp}(z)$, $u \neq v$ with $u \pm v \in 2L^*$ and $z(v) \leq z(u)$, then $(C, \tilde{z})$ with*

$$\tilde{z}(t) = \begin{cases} z(u) - z(v), & \text{if } t = \pm u \\ 0, & \text{if } t = \pm v, \\ 2z(v) + z(t), & \text{if } t \in \{\frac{\pm u \pm v}{2}\}, \\ z(t), & \text{otherwise,} \end{cases}$$

   is a solution for (7.11).

2. *If there is $u \in \mathrm{supp}(z)$ and $u = kv$ for some integer $k \geq 2$, then $(C, \tilde{z})$ with*

$$\tilde{z}(t) = \begin{cases} 0, & \text{if } t = \pm u, \\ z(v) + k^2 z(u), & \text{if } t = \pm v, \\ z(t), & \text{otherwise,} \end{cases}$$

   is a solution for (7.11).

In both cases, $\tilde{z}$ satisfies

$$\sum_{t \in L^*} z(t) t t^\top = \sum_{t \in L^*} \tilde{z}(t) t t^\top \quad \text{and} \quad \sum_{t \in L^*} z(t) < \sum_{t \in L^*} \tilde{z}(t).$$

Proof. (1) By construction, we have

$$\sum_{t \in L^*} z(t) < \sum_{t \in L^*} z(t) + 4z(v) = \sum_{t \in L^*} \tilde{z}(t).$$

Computing

$$\begin{aligned}&z(u)uu^\top + z(v)vv^\top = (z(u)-z(v))uu^\top + z(v)(uu^\top + vv^\top)\\ &= (z(u)-z(v))uu^\top + 2z(v)\left(\left(\frac{u+v}{2}\right)\left(\frac{u+v}{2}\right)^\top + \left(\frac{u-v}{2}\right)\left(\frac{u-v}{2}\right)^\top\right),\end{aligned}$$

(and analogously for the pair $-u, -v$) we obtain $\sum_{t\in L^*} z(t)tt^\top = \sum_{t\in L^*} \tilde{z}(t)tt^\top$ and $CI - 4\pi^2 \sum_{t\in L^*} \tilde{z}(t)tt^\top \in \mathcal{S}_n^+$.

Moreover, by the subquadratic inequality,

$$\begin{aligned}&1-\cos(2\pi u^\top x) + 1 - \cos(2\pi v^\top x)\\ &= 1-\cos\left(2\pi\left(\frac{u+v}{2}+\frac{u-v}{2}\right)^\top x\right) + 1 - \cos\left(2\pi\left(\frac{u+v}{2}-\frac{u-v}{2}\right)^\top x\right)\\ &\le 2\left(1-\cos\left(2\pi\left(\frac{u+v}{2}\right)^\top x\right)\right) + 2\left(1-\cos\left(2\pi\left(\frac{u-v}{2}\right)^\top x\right)\right).\end{aligned}$$

Thus, for every $x \in V(L)$

$$|x|^2 \le 2\sum_{t\in L^*} z(t)(1-\cos(2\pi t^\mathsf{T} x)) \le 2\sum_{t\in L^*} \tilde{z}(t)(1-\cos(2\pi t^\mathsf{T} x)),$$

implying that $(C, \tilde{z})$ is feasible for (7.11) with the desired properties.

(2) The proof is analogous to (1). □

The lemma gives rise to an algorithmic way to transform a feasible solution (C, z) towards a solution $(C, \tilde{z})$ such that $\tilde{z}$ has only support on at most one primitive lattice element per coset $u + 2L^*$. Roughly speaking, start with any solution and apply the above lemma "as long as possible", i.e. as long as there are pairs of vectors that satisfy *(1)* or *(2)* of the above lemma.

Theorem 7.4.2. *For any n-dimensional lattice L, the torus $\mathbb{R}^n/L$ has a finite-dimensional least Euclidean distortion embedding. In particular, the program* (7.11) *has an optimal solution (C, z) such that*

1. $|\operatorname{supp}(\tilde{z}) \cap (v + 2L^*)| \le 1$ *for every coset $v + 2L^*$ of $L^*/2L^*$.*
2. *Every $u \in \operatorname{supp}(z)$ is primitive in L^*.*

Note that claim *(1)* shows that there are at most $2^n - 1$ non-zero elements in the support of z, therefore we obtain an embedding into a space of dimension at most $2^n - 1$.

Proof. As a consequence of Remark 7.2.2, there is an optimal solution (C, z_0) with $z_0 \in \ell_1(L^*)$ for (7.11). Our goal is to construct a sequence of solutions (C, z_m) for (7.11) that converges to a solution that satisfies *(1)* and *(2)*. Let

$$A_z = \{\{u, v\} : u \neq -v, u \pm v \in 2L^*, u, v \in \operatorname{supp}(z)\} \tag{7.19}$$

and let $(z_m)_m$ be a sequence where z_m is obtained from z_{m-1} by applying transformation *(1)* of Lemma 7.4.1 to an arbitrary pair $\{u, v\} \in A_z$ (the actual choice of the pair does not matter). Due to Lemma 7.4.1, the pair (C, z_m) is feasible for (7.11) and we have

$$\sum_{u \in L^*} z_m(u) u u^\top = \sum_{u \in L^*} z_{m+1}(u) u u^\top \quad \text{and} \quad Z_m < Z_{m+1} \quad \text{for all } m \in \mathbb{N},$$

where $Z_m := \sum_{u \in L^*} z_m(u)$. The sequence Z_m is monotonously increasing but bounded since $CI - 4\pi^2 \sum_{u \in L^*} z_m(u) u u^\top \in \mathcal{S}_n^+$ enforces that

$$\begin{aligned} 0 \leq \operatorname{Tr}\Big(CI - 4\pi^2 \sum_{u \in L^*} z_m(u) u u^\top\Big) &= Cn - 4\pi^2 \sum_{u \in L^*} z_m(u)|u|^2 \\ &\leq Cn - 4\pi^2 \lambda(L^*) \sum_{u \in L^*} z_m(u). \end{aligned}$$

Hence, by monotone convergence, the sequence Z_m converges.

Now we claim that $\lim_{m\to\infty} z_m(u)$ exists for all $u \in L^*$. Therefore, assume that $\{u_m, v_m\} \in A_{z_{m-1}}$ is chosen in the iteration from z_{m-1} to z_m. Assume that $z_{m-1}(u_m) \geq z_{m-1}(v_m)$. Then, using Lemma 7.4.1, we obtain

$$\sum_{u \in L^*} |z_m(u) - z_{m-1}(u)| = 3 \cdot 4 z_{m-1}(v_m) = 3(Z_m - Z_{m-1}).$$

The right hand side converges to zero, therefore the sequence z_m converges pointwise, i.e. there is $z \in \ell_1(L^*)$ such that

$$\lim_{m\to\infty} z_m(u) = z(u) \quad \text{for all } u \in L^*.$$

Next, we will show that z satisfies *(1)*, which is equivalent to $A_z = \emptyset$. But this simply follows by construction: For every pair $\{u, v\} \in A_{z_m}$ we have

$$\lim_{m\to\infty} \min\{z_m(u), z_m(v)\} = 0.$$

This holds because if there was $\{u, v\} \in A_z$ and $\varepsilon > 0$ such that for all M there was $m \geq M$ with $\min\{z_m(u), z_m(v)\} \geq \varepsilon$, then according to the construction of z_m there would also be $m' \geq m$ with

$$Z_{m'} - Z_m \geq 4 \min\{z_m(u), z_m(v)\} \geq 4\varepsilon.$$

This would be a contradiction to the convergence of the sequence Z_m.

Now, if there is $u \in \operatorname{supp}(z)$ with $u = kv$ for some $k \geq 2$, we may apply *(2)* of Lemma 7.4.1 to obtain a new feasible solution $\tilde{z}$ with $v \in \operatorname{supp}(\tilde{z})$ and $z(u) = 0$. This solution may contain a pair $(u, v) \in A_{\tilde{z}}$. But in this case, we may again apply *(1)* of Lemma 7.4.1.

By continuing like this, we will finally end up with a solution that has only support on primitive vectors and on one vector per coset, thus satisfying properties *(1)* and *(2)*. □

Unfortunately, the proof does not give a bound on $\max\{|u| : u \in \operatorname{supp}(\tilde{z})\}$ for $\tilde{z}$ constructed in Theorem 7.4.2.

7.5 Improved lower bound

In this section we apply Theorem 7.2.5 to get a constant factor improvement over (7.4), basically without any effort.

Theorem 7.5.1. *Let L be an n-dimensional lattice, then*

$$c_2(\mathbb{R}^n/L) \geq \frac{\pi\lambda(L^*)\mu(L)}{\sqrt{n}}.$$

Proof. Let y be a vertex of the Voronoi cell $V(L)$ which realizes the covering radius, that is $|y| = \mu(L)$ and so y is a "deep hole" of L. Choose $\nu = \frac{\lambda(L^*)^2}{2n}\delta_y$ to be a point measure supported at y and set $Y = \frac{1}{n}I$. Then (ν, Y) is feasible for (7.12) because

$$\int_{V(L)} (1-\cos(2\pi u^\mathsf{T} x))\, d\nu(x) = (1-\cos(2\pi u^\mathsf{T} y))\frac{\lambda(L^*)^2}{2n} \leq \frac{\lambda(L^*)^2}{n} \leq \frac{|u|^2}{n} = u^\mathsf{T} Y u$$

for every $u \in L^* \setminus \{0\}$. Hence, by Theorem 7.2.5,

$$c_2(\mathbb{R}^n/L)^2 \geq 2\pi^2 \int_{V(L)} |x|^2 \, d\nu(x) = \frac{\pi^2 \lambda(L^*)^2 \mu(L)^2}{n}. \qquad \square$$

7.6 Least distortion embeddings of $\mathbb{R}^n/\mathbb{Z}^n$ and of orthogonal decompositions

As our second application of Theorem 7.2.5, through Theorem 7.5.1, we prove that the standard embedding (7.2) of the standard torus is indeed a least distortion embedding. It is somewhat surprising that this result is new. We also note that one can easily use the same argument to capture the case of flat tori whose lattices have an orthogonal basis.

Theorem 7.6.1. *The standard embedding $\varphi : \mathbb{R}^n/\mathbb{Z}^n \to \mathbb{R}^{2n}$ of the standard torus $\mathbb{R}^n/\mathbb{Z}^n$ given by*

$$\varphi(x_1, \ldots, x_n) = (\cos 2\pi x_1, \sin 2\pi x_1, \ldots, \cos 2\pi x_n, \sin 2\pi x_n)$$

is a least distortion embedding with distortion $c_2(\mathbb{R}^n/\mathbb{Z}^n) = \pi/2$.

Proof. We have $\lambda(\mathbb{Z}^n) = 1$ and $\mu(\mathbb{Z}^n) = \sqrt{n/4}$, so $c_2(\mathbb{R}^n/\mathbb{Z}^n) \geq \pi/2$ by Theorem 7.5.1.

To show the corresponding upper bound we show that the embedding

$$\phi(x_1, \ldots, x_n) = \frac{1}{\sqrt{32}}(e^{-2\pi i x_1}, \ldots, e^{-2\pi i x_n})$$

has contraction 1 and expansion $\pi/2$. Then turning ϕ into a real embedding and rescaling does not change the distortion and gives the standard embedding φ.

To show that ϕ has contraction 1 and expansion $\pi/2$ it suffices to prove that $(\frac{\pi^2}{4}, z)$ with

$$z(u) = \begin{cases} \frac{1}{32} & \text{if } u = \pm e_i, \\ 0 & \text{otherwise,} \end{cases}$$

is a feasible solution for (7.11).

The expansion equals $\pi/2$ because

$$CI - 4\pi^2 \sum_{u\in\mathbb{Z}^n} z(u)uu^{\mathsf{T}} = \frac{\pi^2}{4}I - 4\pi^2\frac{2}{32}I = 0.$$

Moreover, to show that the contraction equals 1 we need to verify the inequality

$$\sum_{i=1}^{n} x_i^2 \leq \frac{4}{32}\sum_{i=1}^{n}(1-\cos(2\pi x_i)) \quad \text{for all } x \in V(\mathbb{Z}^n) = [-1/2, 1/2]^n, \tag{7.20}$$

which we check summand by summand, that is $x_i^2 \leq \frac{1}{8}(1-\cos(2\pi x_i))$, where we have equality if $x_i = \pm 1/2$. To do so we show that the logarithm of the quotient

$$x_i \mapsto \log\left(\frac{x_i^2}{1-\cos(2\pi x_i)}\right)$$

is convex on $(-1/2, 1/2)$ which follows by taking the second derivative: For $x_i \in (0, 1/2)$ we have

$$\frac{\partial^2}{\partial x_i^2}\log\left(\frac{x_i^2}{1-\cos(2\pi x_i)}\right) = -\frac{2}{x_i^2} + \frac{4\pi^2}{1-\cos(2\pi x_i)} \geq 0,$$

where we used the inequality $1-\cos(2\pi x_i) \leq 2\pi^2 x_i^2$. □

Here it is interesting to note that even for the rather trivial standard embedding of the standard torus the structure of the most contracted pairs is rich. Every center of every face of the Voronoi cell $V(\mathbb{Z}^n)$ gives a most contracted pair, see Figure 7.1a.

Recapulating the above proof, one recognizes that at its heart is the verification of inequality (7.20). Here one reduces the situation from $\mathbb{Z}^n$ to $\mathbb{Z}$. This works because $\mathbb{Z}^n$ can be orthogonally decomposed as the direct sum of n copies of $\mathbb{Z}$ and this can be done in generality as the following theorem demonstrates.

Theorem 7.6.2. *Let $L \subseteq \mathbb{R}^n$ be a lattice such that L decomposes as the orthogonal direct sum of lattices $L_1, \ldots, L_m$, i.e.*

$$L = L_1 \perp L_2 \perp \ldots \perp L_m.$$

Then

$$c_2(\mathbb{R}^n/L) = \max\{c_2(\mathbb{R}^{n_j}/L_j) : j = 1, \ldots, m\},$$

where $\mathbb{R}^{n_j}$ is (isometric) to the Euclidean space spanned by L_j.

Proof. Any Euclidean embedding of $\mathbb{R}^n/L$ gives a Euclidean embedding of $\mathbb{R}^{n_j}/L_j$ which immediately gives the inequality $c_2(\mathbb{R}^n/L) \geq \max\{c_2(\mathbb{R}^{n_j}/L_j) : j = 1, \ldots, m\}$.

Also the reverse inequality is easy to see. Let $\varphi_j : \mathbb{R}^{n_j}/L_j \to H_j$ be a Euclidean embedding of $\mathbb{R}^{n_j}/L_j$ with distortion C_j scaled so that the contraction is 1 and the expansion is C_j. We identify $\mathbb{R}^n/L \cong \mathbb{R}^{n_1}/L_1 \perp \ldots \perp \mathbb{R}^{n_m}/L_m$, write $x \in \mathbb{R}^n/L$ as $x = (x_1, \ldots, x_m)$ with $x_j \in \mathbb{R}^{n_j}/L_j$ so that $d_{\mathbb{R}^n/L}(x, y)^2 = \sum_{j=1}^m d_{\mathbb{R}^{n_j}/L_j}(x_j, y_j)^2$. Then

$$\varphi : \mathbb{R}^n/L \to H := H_1 \perp \ldots \perp H_m, \quad (x_1, \ldots, x_m)+L \mapsto (\varphi_1(x_1), \ldots, \varphi_m(x_m))$$

is a Euclidean embedding of $\mathbb{R}^n/L$ into the Hilbert space H. Its distortion is at most $\max\{C_j : j = 1, \ldots, m\}$ because for every pair $x, y \in \mathbb{R}^n/L$

$$\begin{aligned}|\varphi(x) - \varphi(y)|^2 &= \sum_{j=1}^m |\varphi_j(x_j) - \varphi_j(y_j)|^2 \leq \sum_{j=1}^m C_j^2 d_{\mathbb{R}^{n_j}/L_j}(x_j, y_j)^2 \\ &\leq \max_{j=1,\ldots,m} C_j^2 \sum_{j=1}^m d_{\mathbb{R}^{n_j}/L_j}(x_j, y_j)^2 = \max_{j=1,\ldots,m} C_j^2 \, d_{\mathbb{R}^n/L}(x, y)^2,\end{aligned}$$

showing that the expansion of φ at most $\max\{C_j : j = 1, \ldots, m\}$ and, in exactly the same way, one shows that the contraction of φ is at most 1. □

7.7 Least distortion embeddings of two-dimensional flat tori

In this section we will construct least Euclidean distortion embeddings of flat tori in dimension 2.

First, as a simple corollary of Theorem 7.2.4, we will give a recipe to construct (possibly non-optimal) embeddings of flat tori of arbitrary

dimension provided that they satisfy the following assumption:

$$\text{There exist } u_1, \dots, u_k \in L^*,\ z_1, \dots, z_k \geq 0 \text{ such that } 4\pi^2 \sum_{i=1}^{k} z_i u_i u_i^{\mathsf{T}} = I. \tag{7.21}$$

As we will prove in Lemma 7.7.2, condition (7.21) can be realized for every 2-dimensional lattice, Another example for lattices that satisfy condition (7.21) are duals of root lattices. If L^* is a root lattice, then assumption (7.21) is satisfied. In this case the root system $R \subseteq L^*$ of L^* forms a spherical 2-design, implying that

$$4\pi^2 \alpha \sum_{u \in R} u u^{\mathsf{T}} = I$$

for some positive constant α. We refer to the monograph by Venkov [Ven01] for more information on root lattices and spherical designs.

Corollary 7.7.1. *Let $L \subseteq \mathbb{R}^n$ be a lattice that satisfies* (7.21)*. Then*

$$\varphi : \mathbb{R}^n/L \to \mathbb{C}^k, \quad \varphi(x) = (\sqrt{Dz_1} e^{2i\pi u_1^{\mathsf{T}} x}, \dots, \sqrt{Dz_k} e^{2i\pi u_k^{\mathsf{T}} x})$$

with

$$D = \max \left\{ \frac{|x|^2}{2 \sum_{i=1}^{k} z_i (1 - \cos(2\pi u_i^{\mathsf{T}} x))} : x \in V(L) \setminus \{0\} \right\} \tag{7.22}$$

is a Euclidean embedding of $\mathbb{R}^n/L$ with distortion $\sqrt{D}$. In particular,

$$c_2(\mathbb{R}^n/L)^2 \leq D. \tag{7.23}$$

Proof. The pair $((Dz_i)_{1 \leq i \leq k}, D)$ is a feasible solution for the primal optimization problem (7.11). □

Except for the easiest case of the standard torus we do not know how to determine D explicitly. Unfortunately, it seems to be difficult to compute most contracted pairs $(0, x)$, i.e. vectors $x \in V(L)$ that are maximizers of the right hand side of (7.22).

Next, we show that Corollary 7.7.1 can be applied to every 2-dimensional lattice. For this we will use the concept of an *obtuse*

superbasis. An obtuse superbasis of an n-dimensional lattice L is a basis $u_1, \ldots, u_n$ of L enlarged by the vector $u_0 = -u_1 - \cdots - u_n$ so that these $n+1$ vectors pairwise form non-acute angles, i.e.

$$u_i^\mathsf{T} u_j \leq 0 \text{ for all } 0 \leq i < j \leq n. \tag{7.24}$$

It is known that up to dimension 3 all lattices have an obtuse superbasis, but from dimension 4 on this is no longer the case, see for instance [CS92].

Lemma 7.7.2. *If L is a two-dimensional lattice, then its dual lattice L^* satisfies* (7.21).

Proof. Let u_0, u_1, u_2 be an obtuse superbasis of L^*. We will show that there are non-negative coefficients z_0, z_1, z_2 such that

$$I = z_0 u_0 u_0^\mathsf{T} + z_1 u_1 u_1^\mathsf{T} + z_2 u_2 u_2^\mathsf{T},$$

and therefore condition (7.21) holds.

We may assume that $|u_1| \geq |u_0| = 1$, by scaling and renumbering. Then, by Gram-Schmidt orthogonalization, u_0 and $w_1 := u_1 - (u_0^\mathsf{T} u_1) u_0$ are orthogonal and so

$$\begin{aligned} I &= u_0 u_0^\mathsf{T} + \frac{1}{|w_1|^2} w_1 w_1^\mathsf{T} \\ &= \left(1 + \frac{(u_0^\mathsf{T} u_1)^2}{|w_1|^2}\right) u_0 u_0^\mathsf{T} + \frac{1}{|w_1|^2} u_1 u_1^\mathsf{T} - \frac{u_0^\mathsf{T} u_1}{|w_1|^2}(u_0 u_1^\mathsf{T} + u_1 u_0^\mathsf{T}). \end{aligned}$$

Using

$$u_2 u_2^\mathsf{T} = (-u_0 - u_1)(-u_0 - u_1)^\mathsf{T} = u_0 u_0 + u_1 u_0^\mathsf{T} + u_0 u_1^\mathsf{T} + u_1 u_1^\mathsf{T},$$

yields

$$I = \left(1 + \frac{(u_0^\mathsf{T} u_1)^2 + u_0^\mathsf{T} u_1}{|w_1|^2}\right) u_0 u_0^\mathsf{T} + \frac{1 + u_0^\mathsf{T} u_1}{|w_1|^2} u_1 u_1^\mathsf{T} - \frac{u_0^\mathsf{T} u_1}{|w_1|^2} u_2 u_2^\mathsf{T}.$$

To prove that the three coefficients in the above sum are non-negative, observe for the third coefficient that $u_0^\mathsf{T} u_1 \leq 0$. For the second coefficient

$$0 \leq -u_0^\mathsf{T} u_2 = u_0^\mathsf{T} u_0 + u_0^\mathsf{T} u_1 = 1 + u_0^\mathsf{T} u_1.$$

For the first coefficient we compute the squared norm $|w_1|^2 = |u_1|^2 - (u_0^\mathsf{T} u_1)^2$ and see that the first coefficient is nonnegative if and only if $|u_1|^2 \geq -u_0^\mathsf{T} u_1$, which is true because $|u_1| \geq |u_0| = 1$ by assumption. □

Now, to verify that the embedding of Corollary 7.7.1 is indeed a least Euclidean distortion embedding for two-dimensional flat tori, we construct a dual solution for (7.12) that shows that the upper bound (7.23) is sharp.

Theorem 7.7.3. *Let $L \subseteq \mathbb{R}^2$ be a 2-dimensional lattice, then $c_2(\mathbb{R}^2/L)^2 = D$, where D is defined in* (7.22).

Proof. Let u_0, u_1, u_2 be an obtuse superbasis of L^*. We may assume, see for example [CS92], that this superbasis is chosen in a way such that u_i is a shortest vector in its coset $u_i + 2L^*$, with $i = 0, 1, 2$. By Lemma 7.7.2 we can determine coefficients $z_0, z_1, z_2 \geq 0$ such that $4\pi^2 \sum_{i=0}^2 z_i u_i u_i^\mathsf{T} = I$.

Furthermore, let $\bar{x} \in V(L)$ be a vector such that $(0, \bar{x})$ is a most contracted pair for the embedding φ of Corollary 7.7.1, that is, $\bar{x}$ is a maximizer for (7.22).

We define the pair (Y, ν) as follows: Set $\beta = \frac{D}{2\pi^2 |\bar{x}|^2}$, define Y via

$$\operatorname{Tr}(u_i u_i^\mathsf{T} Y) = \beta(1 - \cos(2\pi u_i^\mathsf{T} \bar{x})), \quad i \in \{0, 1, 2\}, \tag{7.25}$$

and let $\nu = \beta \delta_{\bar{x}}$ be a point measure supported only on $\bar{x}$. We now verify that this pair is a feasible dual solution with objective value D.

We have

$$\operatorname{Tr}(Y) = \operatorname{Tr}(YI) = 4\pi^2 \sum_i z_i \operatorname{Tr}(u_i u_i^\mathsf{T} Y) = 4\pi^2 \beta \sum_i z_i (1 - \cos(2\pi u_i^\mathsf{T} \bar{x})) = 1.$$

Equation (7.25) together with Lemma 7.3.3 implies

$$\operatorname{Tr}(u u^\mathsf{T} Y) \geq \beta(1 - \cos(2\pi u^\mathsf{T} \bar{x})) \quad \text{for all } u \in L^*.$$

Finally, it remains to show that Y is positive semidefinite. For this we compute its Gram matrix B with respect to u_0, u_1, that is

$$B_{ij} = u_i^\mathsf{T} Y u_j = \frac{1}{2} \operatorname{Tr}((u_i u_j^\mathsf{T} + u_j u_i^\mathsf{T}) Y), \quad 0 \leq i, j \leq 1.$$

Then $B_{ii} = \mathrm{Tr}(u_i u_i^\mathsf{T} Y) = 2\beta \sin^2(\pi u_i^\mathsf{T} \bar{x})$. Since

$$u_0 u_1^\mathsf{T} + u_1 u_0^\mathsf{T} = u_2 u_2^\mathsf{T} - u_0 u_0^\mathsf{T} - u_1 u_1^\mathsf{T},$$

we get

$$\begin{aligned} B_{01} &= \frac{\beta}{2}\Big(2\sin^2(\pi u_2^\mathsf{T}\bar{x}) - 2\sin^2(\pi u_1^\mathsf{T}\bar{x}) - 2\sin^2(\pi u_0^\mathsf{T}\bar{x})\Big) \\ &= 2\beta \sin(\pi u_0^\mathsf{T}\bar{x}) \sin(\pi u_1^\mathsf{T}\bar{x}) \cos(\pi(u_0+u_1)^\mathsf{T}\bar{x}). \end{aligned}$$

From this we see that matrix B is the Schur-Hadamard (entry-wise) product of the positive semidefinite rank-one matrix xx^T with $x_i = \sqrt{2\beta}\sin(\pi u_i^\mathsf{T}\bar{x})$ and the symmetric matrix $M \in \mathbb{R}^{2\times 2}$ defined by

$$M_{ij} = \begin{cases} 1 & \text{if } i = j \\ \cos(\pi(u_0+u_1)^\mathsf{T}\bar{x}) & \text{if } (i,j) \in \{(0,1),(1,0)\}. \end{cases}$$

The matrix M is positive semidefinite because $M_{ii} \geq 0$ and

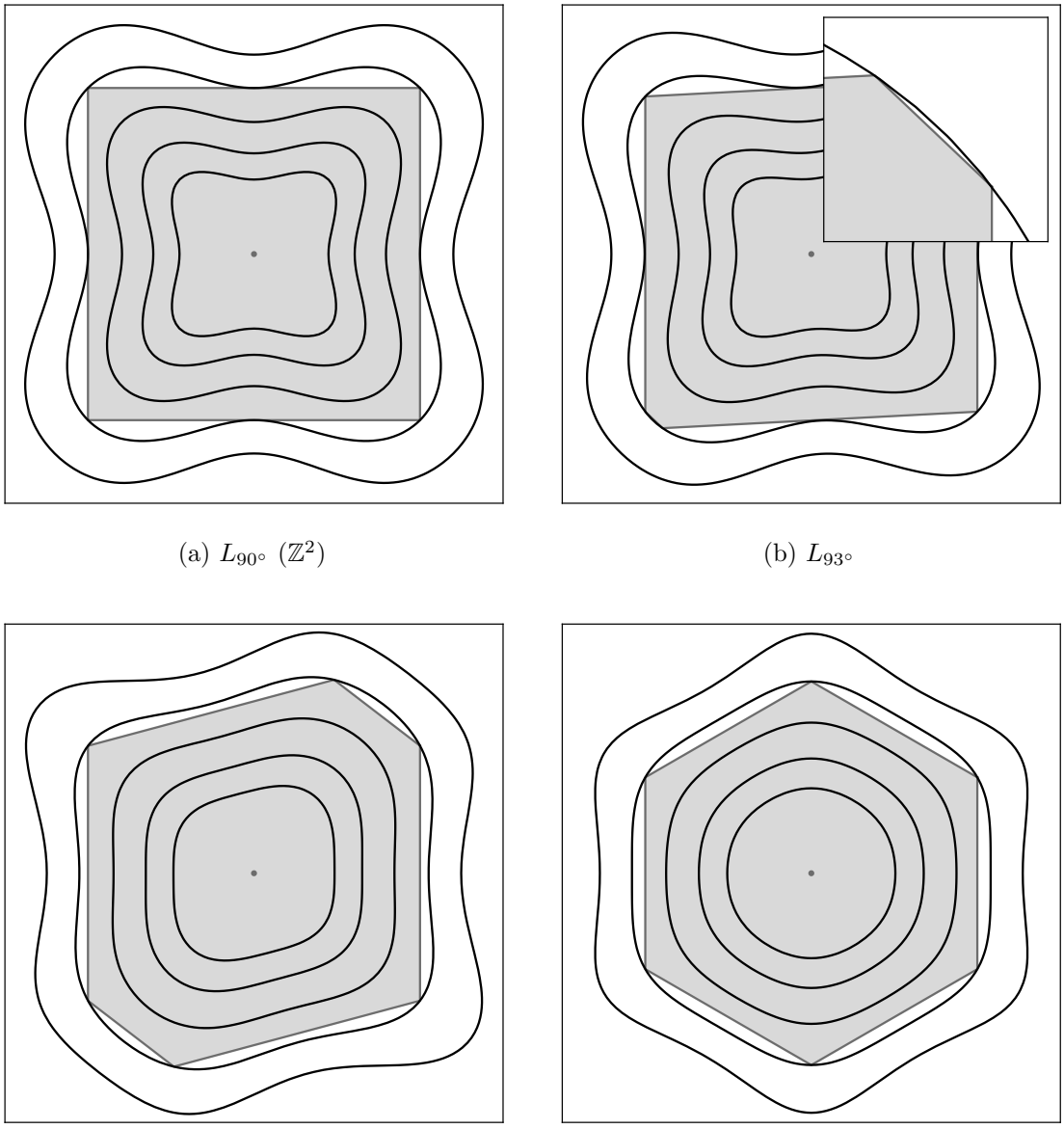
$$\det(M) = 1 - \cos^2(\pi(u_0+u_1)^\mathsf{T}\bar{x}) \geq 0.$$

Thus B, and therefore also Y, is positive semidefinite, which finishes the proof. □

To conclude the discussion of 2-dimensional lattices Figure 7.1 collects an illustration of the behavior of the distortion function defined in (7.22) and the most contracted pairs, applying the above results, for a selection of 2-dimensional lattices.

7.8 Discussion and open questions

In this paper we derived an infinite-dimensional semidefinite program to determine the least distortion Euclidean embedding of a flat torus. It would be very interesting to show that this infinite-dimensional semidefinite program can in fact be turned into a finite-dimensional semidefinite program. Then one could, similarly to the case of finite metric spaces, algorithmically determine least distortion Euclidean embeddings of flat tori; at least up to any desired precision.



(a) L_{90° ($\mathbb{Z}^2$)

(b) L_{93°

(c) L_{105°

(d) L_{120° (A_2^*)

Figure 7.1: Let L_φ be the lattice spanned by $v_1 = e_1$, $v_2 = R_\varphi e_1$, where R_φ is the rotation by φ degrees (counter-clockwise).
The Voronoi cell of any lattice in $\mathbb{R}^2$ is either a rectangle or a hexagon. We plot contour lines of the distortion function for a selection of lattices that illustrate how the distortion function and the most contracted pairs vary with the shape of the Voronoi cell. L_{93° shows what happens for almost degenerated hexagons, i.e. lattices close to the standard lattice. A zoom into the behavior around the short edge of the hexagon is included to illustrate that the most contracted points are all vertices of the Voronoi-cell.

For this a characterization of the most contracted pairs is needed. We believe that the most contracted pairs are always of the form $(0, y)$ and y is a center of a face of the Voronoi cell. However, we do not know whether such a y can only lie on the Voronoi cell's boundary. We do not even know whether there are only finitely many most contracted pairs.

We also do not know how to restrict the variable $z \in \ell_1(L^*)$ to finite dimension, even though Theorem 7.4.2 shows that we can always find a finite-dimensional least distortion embedding. Obtaining a bound on the maximally needed length of a support vector in the cosets $L^*/2L^*$ would solve this problem.

Another interesting problem is to determine n-dimensional lattices which maximize the distortion among all n-dimensional lattices.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie agreement No 764759. F.V. is partially supported by the SFB/TRR 191 "Symplectic Structures in Geometry, Algebra and Dynamics", F.V. and M.C.Z. are partially supported "Spectral bounds in extremal discrete geometry" (project number 414898050), both funded by the DFG. A.H. is partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – Cluster of Excellence Matter and Light for Quantum Computing (ML4Q) EXC 2004/1 – 390534769.

Acknowledgments

First, I would like to thank my two advisors Frank and David. You both have a sometimes different but absolutely inspiring way of conveying passion and ideas in math, computer science and physics. Maybe personally the most important thing I am grateful for: you gave me the freedom to do my own research even if this sometimes lead to dead ends.

Furthermore, I would like to thank the people I worked on projects with and had many insightful discussions with: Markus, Felipe, Valentin, Marc, Aurelio, Antonia, Moritz, Michael, Robert and Cihan.

Not to forget are the people from the discrete math group and the quantum info group in Cologne. Especially, Andreas and Karla for proofreading and open ears for any type of math or non-math discussion.

Last bot not least I would like to thank my old fellow study mates from my Bachelors and Masters in Cologne. You have made the last 9 years an amazing time!

Bibliography

[ADGS18] M. Ahmadi, B. Dang, G. Gour, and B. C. Sanders. Quantification and manipulation of magic states. *Phys. Rev. A*, 97(6):062332, 2018. doi:10.1103/PhysRevA.97.062332.

[ADRSD15] D. Aggarwal, D. Dadush, O. Regev, and N. Stephens-Davidowitz. Solving the shortest vector problem in 2^n time using discrete Gaussian sampling. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing (FOCS)*, pages 733–742. Association for Computing Machinery, New York, NY, USA, 2015. doi:10.1145/2746539.2746606.

[ADSD15] D. Aggarwal, D. Dadush, and N. Stephens-Davidowitz. Solving the closest vector problem in 2^n time – the discrete Gaussian strikes again! In *Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 563–582. IEEE Computer Society, New York, NY, USA, 2015. doi:10.1109/FOCS.2015.41.

[AG03] Farid Alizadeh and Donald Goldfarb. Second-order cone programming. *Math. Program. 95, no. 1, Ser. B*, 2003. doi:10.1007/s10107-002-0339-5.

[AMM85] I. Aharoni, B. Maurey, and B. S. Mityagin. Uniform embeddings of metric spaces and of Banach spaces into Hilbert spaces. *Israel J. Math.*, 52:251–265, 1985. doi:doi.org/10.1007/BF02786521.

[ART20] I. Agarwal, O. Regev, and Y. Tang. Nearly optimal embeddings of flat tori. *APPROX/RANDOM*, 176:43:1–43:14, 2020. doi:10.4230/LIPIcs.APPROX/RANDOM.2020.43.

[Asc00] M. Aschbacher. *Finite Group Theory*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2000. doi:10.1017/CBO9781139175319.

[BB09] E. Bannai and E. Bannai. A survey on spherical designs and algebraic combinatorics on spheres. *European Journal of Combinatorics*, 30(6):1392–1425, 2009. doi:https://doi.org/10.1016/j.ejc.2008.11.007.

[BBC+19] S. Bravyi, D. Browne, P. Calpin, E. Campbell, D. Gosset, and M. Howard. Simulation of quantum circuits by low-rank stabilizer decompositions. *Quantum*, 3:181, 2019. doi:10.22331/q-2019-09-02-181.

[BCHK20] M. Beverland, E. Campbell, M. Howard, and V. Kliuchnikov. Lower bounds on the non-Clifford resources for quantum computations. *Quantum Science and Technology*, 5(3):035009, 2020. doi:10.1088/2058-9565/ab8963.

[BDF+99] C. H. Bennett, D. P. DiVincenzo, C. A. Fuchs, T. Mor, E. Rains, P. W. Shor, J. A. Smolin, and W. K. Wootters. Quantum nonlocality without entanglement. *Phys. Rev. A*, 59(2):1070–1091, 1999. doi:10.1103/PhysRevA.59.1070.

[BG16] S. Bravyi and D. Gosset. Improved classical simulation of quantum circuits dominated by Clifford gates. *Phys. Rev. Lett. 116, 250501*, 2016. arXiv:1601.07601v3, doi:10.1103/PhysRevLett.116.250501.

[BGL22] S. Bravyi, D. Gosset, and Y. Liu. How to simulate quantum measurement without computing marginals.

Phys. Rev. Lett., 128:220503, 2022. doi:10.1103/PhysRevLett.128.220503.

[BGSV12] C. Bachoc, D. C. Gijswijt, A. Schrijver, and F. Vallentin. Invariant semidefinite programs. In *Handbook on semidefinite, conic, and polynomial optimization (M.F. Anjos, J.B. Lasserre, eds.)*, pages 219–269. Springer, 2012. doi:doi.org/10.1007/978-1-4614-0769-0_9.

[BJLT12] V. Borrelli, S. Jabrane, F. Lazarus, and B. Thibert. Flat tori in three-dimensional space and convex integration. *Proceedings of the National Academy of Sciences (PNAS)*, 109:7218–7223, 2012. doi:https://doi.org/10.1073/pnas.1118478109.

[BK05] S. Bravyi and A. Kitaev. Universal quantum computation with ideal Clifford gates and noisy ancillas. *Phys. Rev. A 71, 022316*, 2005. doi:10.1103/PhysRevA.71.022316.

[Bos18] J.-B. Bost. Exposé bourbaki 1151: Réseaux euclidiens séries thêta et pentes (d'après W. Banaszczyk, O. Regev, D. Dadush, N. Stephens-Davidowitz, ...). *Séminaire Bourbaki*, 2018.

[BSS16] S. Bravyi, G. Smith, and J. Smolin. Trading classical and quantum computational resources. *Phys. Rev. X 6, 021043*, 2016. doi:10.1103/PhysRevX.6.021043.

[But72] G. J. Butler. Simultaneous packing and covering in Euclidean space. *Proc. London Math. Soc. 25*, s3-25:721–735, 1972. doi:https://doi.org/10.1112/plms/s3-25.4.721.

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. doi:10.1017/CBO9780511804441.

[BVDB+17] J. Bermejo-Vega, N. Delfosse, D. E. Browne, C. Okay, and R. Raussendorf. Contextuality as a resource for models of quantum computation on qubits. *Phys. Rev. Lett. 119, 120505*, 2017. doi:10.1103/PhysRevLett.119.120505.

[Cam20] E. Campbell. Private communication, 2020.

[Cam21] E. Campbell. Private communication, 2021.

[CB09] E. T. Campbell and D. E. Browne. On the structure of protocols for magic state distillation. In A. Childs and M. Mosca, editors, *Theory of Quantum Computation, Communication, and Cryptography*, pages 20–32. Springer Berlin Heidelberg, 2009. doi:https://doi.org/10.1007/978-3-642-10698-9_3.

[CCL12] E. Chitambar, W. Cui, and H.-K. Lo. Increasing entanglement monotones by separable operations. *Phys. Rev. Lett.*, 108(24):240504, 2012. doi:10.1103/PhysRevLett.108.240504.

[CG19] E. Chitambar and G. Gour. Quantum resource theories. *Reviews of Modern Physics*, 91(2):025001, 2019. doi:10.1103/RevModPhys.91.025001.

[CGG+06] C. Cormick, E. F. Galvao, D. Gottesman, J. P. Paz, and A. O. Pittenger. Classicality in discrete wigner functions. *Phys. Rev. A*, 73:012301, 2006. doi:10.1103/PhysRevA.73.012301.

[CGIK21] S. M. Cioabă, H. Gupta, F. Ihringer, and H. Kurihara. *The least Euclidean distortion constant of a distance-regular graph.* 2021. arXiv:2109.09708.

[Chi] E. Chitambar. Local quantum transformations requiring infinite rounds of classical communication. 107(19):190502. doi:10.1103/PhysRevLett.107.190502.

[CKM+16] H. Cohn, A. Kumar, S. Miller, D. Radchenko, and M. Viazovska. The sphere packing problem in dimension 24. *Annals of Mathematics*, 185:1017–1033, 2016. doi:10.4007/annals.2017.185.3.8.

[CKM+22] H. Cohn, A. Kumar, S. D. Miller, D. Radchenko, and M. Viazovska. Universal optimality of the E_8 and Leech lattices and interpolation formulas. *Annals of Mathematics*, 196:983–1082, 2022. doi:https://doi.org/10.4007/annals.2022.196.3.3.

[CLM+14] E. Chitambar, D. Leung, L. Mančinska, M. Ozols, and A. Winter. Everything you always wanted to know about LOCC (but were afraid to ask). *Communications in Mathematical Physics*, 328(1):303–326, 2014. doi:10.1007/s00220-014-1953-9.

[Cou06] Renaud Coulangeon. Spherical designs and zeta functions of lattices. *Int. Math. Res. Not. IMRN*, 49620, 2006. doi:10.1155/IMRN/2006/49620.

[CS88] J. H. Conway and N. J. A. Sloane. Sphere packings, lattices, and groups. *Springer*, 1988. doi:https://doi.org/10.1007/978-1-4757-6568-7.

[CS92] J. H. Conway and N. J. A. Sloane. Low-dimensional lattices VI: Voronoi reduction of three-dimensional lattices. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 436:55–68, 1992. doi:https://doi.org/10.1098/rspa.1992.0004.

[CS12] R. Coulangeon and A. Schürmann. Energy minimization, periodic sets and spherical designs. *Int. Math. Res. Not. IMRN*, pages 829–848, 2012. doi:https://doi.org/10.1093/imrn/rnr048.

[DDM03] J. Dehaene and B. De Moor. The Clifford group, stabilizer states, and linear and quadratic operations over

GF(2). *Phys. Rev. A*, 68(4), 2003. doi:10.1103/PhysRevA.68.042318.

[DFXY09] R. Duan, Y. Feng, Y. Xin, and M. Ying. Distinguishability of Quantum States by Separable Operations. *IEEE Transactions on Information Theory*, 55(3):1320–1330, 2009. doi:10.1109/TIT.2008.2011524.

[DOBV+17] N. Delfosse, C. Okay, J. Bermejo-Vega, D. E. Browne, and R. Raussendorf. Equivalence between contextuality and negativity of the Wigner function for qudits. *New. J. Phys.*, 19(12):123024, 2017. doi:10.1088/1367-2630/aa8fe3.

[Ebe94] W. Ebeling. *Lattices and Codes*. Vieweg, 1994. doi:https://doi.org/10.1007/978-3-658-00360-9.

[EG15] J. Epstein and D. Gottesman. Stabilizer quantum mechanics and magic state distillation, 2015. Masters Essay, Perimeter Institute. URL: http://jeffreymepstein.com/PSIEssay.pdf.

[ERS22] Y. Eisenberg, O. Regev, and N. Stephens-Davidowitz. A tight reverse Minkowski inequality for the Epstein zeta function. *Proceedings of the AMS*, 2022. arXiv:2201.05201.

[FCY+04] D. Fattal, T. S. Cubitt, Y. Yamamoto, S. Bravyi, and I. L. Chuang. Entanglement in the stabilizer formalism. 2004. arXiv:quant-ph/0406168.

[FH91] W. Fulton and J. W. Harris. *Representation Theory: A. First Course*. Springer, 1991. doi:doi.org/10.1007/978-1-4612-0979-9.

[FL20] K. Fang and Z.-W. Liu. No-go theorems for quantum resource purification. *Phys. Rev. Lett.*, 125(6):060405, 2020. doi:10.1103/PhysRevLett.125.060405.

[FM04] Uriel Feige and Daniele Micciancio. The inapproximability of lattice and coding problems with preprocessing. *Journal of Computer and System Sciences*, 69(1):45–67, 2004. Special Issue on Computational Complexity 2002. doi:https://doi.org/10.1016/j.jcss.2004.01.002.

[Fol95] G. B. Folland. *A course in abstract harmonic analysis*. CRC Press, 1995. doi:https://doi.org/10.1201/b19172.

[FR13] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer, 2013. doi:10.1007/978-0-8176-4948-7.

[Fra51] J. S. Frame. The classes and representations of the groups of 27 lines and 28 bitangents. *Ann. Mat. Pura Appl. (4) 32*, 4:83–119, 1951. doi:doi.org/10.1007/BF02417955.

[Fra70] J.S. Frame. The characters of the Weyl group E_8. In *Computational Problems in Abstract Algebra*, pages 111–130. Pergamon, 1970. doi:https://doi.org/10.1016/B978-0-08-012975-4.50017-5.

[FRB18] M. Frembs, S. Roberts, and S. D. Bartlett. Contextuality as a resource for measurement-based quantum computation beyond qubits. *New J. Phys. 20, 103011*, 2018. doi:10.1088/1367-2630/aae3ad.

[Gal05] E. F. Galvão. Discrete Wigner functions and quantum computational speedup. *Phys. Rev. A*, 71(4):042302, 2005. doi:10.1103/PhysRevA.71.042302.

[Got97] D. Gottesman. *Stabilizer Codes and Quantum Error Correction*. PhD thesis, California Institute of Technology, 1997. arXiv:9705052.

[Got99] D. Gottesman. The Heisenberg representation of quantum computers. *Group22: Proceedings of the XXII International Colloquium on Group Theoretical Methods in Physics, eds. S. P. Corney, R. Delbourgo, and P. D. Jarvis, pp. 32-43 (Cambridge, MA, International Press)*, 1999. arXiv:quant-ph/9807006v1.

[Gro06] D. Gross. Hudson's theorem for finite-dimensional quantum systems. *J. Math. Phys. 47, 122107*, 2006. doi:10.1063/1.2393152.

[Gro07] D. Gross. Non-negative Wigner functions in prime dimensions. *Applied Physics B*, 86(3):367–370, 2007. doi:10.1007/s00340-006-2510-9.

[HC17a] M. Howard and E. Campbell. Application of a resource theory for magic states to fault-tolerant quantum computing. *Phys. Rev. Lett. 118, 090501*, 2017. doi:10.1103/PhysRevLett.118.090501.

[HC17b] M. Howard and E. T. Campbell. Application of a resource theory for magic states to fault-tolerant quantum computing. *Phys. Rev. Lett. 118, 090501*, 2017. doi:10.1103/PhysRevLett.118.090501.

[Hei19] A. Heimendahl. The stabilizer polytope and contextuality for qubit systems. Master's thesis, University of Cologne, 2019. URL: http://www.mi.uni-koeln.de/opt/wp-content/uploads/2020/07/MT_Arne_Heimendahl.pdf.

[Hei21] M. Heinrich. *On stabiliser techniques and their application to simulation and certification of quantum devices*. PhD thesis, University of Cologne, 2021. URL: https://kups.ub.uni-koeln.de/50465/.

[HG19] M. Heinrich and D. Gross. Robustness of magic and symmetries of the stabiliser polytope. *Quantum*, 3:132, 2019. doi:10.22331/q-2019-04-08-132.

[HMMVG21] A. Heimendahl, F. Montealegre-Mora, F. Vallentin, and D. Gross. Stabilizer extent is not multiplicative. *Quantum*, 5:400, 2021. doi:10.22331/q-2021-02-24-400.

[How73] R. Howe. Invariant theory and duality for classical groups over finite fields with applications to their singular representation theory. preprint, Yale University, 1973. URL: https://cpb-us-w2.wpmucdn.com/blog.nus.edu.sg/dist/3/12136/files/2021/02/Howe-Invariant-Theory-and-Duality-finite-fields.pdf.

[How88] R. Howe. The oscillator semigroup. In *The Mathematical Heritage of Hermann Weyl*, volume 48 of *Proc. Sympos. Pure Math.*, pages 61–132. Amer. Math. Soc., 1988.

[HR13] I. Haviv and O. Regev. The Euclidean distortion of flat tori. *J. Topol. Anal.*, 5(2013):205–223, 2013. doi:doi.org/10.1007/978-3-642-15369-3_18.

[HWVE14] M. Howard, J. Wallman, V. Veitch, and J. Emerson. Contextuality supplies the 'magic' for quantum computation. *Nature*, 510(7505):351–355, 2014. doi:10.1038/nature13460.

[JR11] P. Jenkins and J. Rouse. Bounds for coefficients of cusp forms and extremal lattices. *Bull. London Math. Soc.*, 43:927–938, 2011. doi:https://doi.org/10.1112/blms/bdr030.

[Ker94] M. Kervaire. Unimodular lattices with a complete root system. *L'Enseign. Math.*, 40:59–140, 1994.

[KG15] R. Kueng and D. Gross. Qubit stabilizer states are complex projective 3-designs. 2015. arXiv:1510.02767v1.

[Kin03] O. D. King. A mass formula for unimodular lattices with no roots. *Math. Comput.*, 72(242):839–863, 2003. doi:10.1090/S0025-5718-02-01455-2.

[KK07] M. Koecher and A. Krieg. *Elliptische Funktionen und Modulformen*. Springer, 2007. doi:10.1007/978-3-540-49325-9.

[KK15] T. Kobayashi and T. Kondo. The Euclidean distortion of generalized polygons. *Adv. Geom.*, 15:499–506, 2015. doi:https://doi.org/10.1515/advgeom-2015-0023.

[KL19] W. M. Kirby and P. J. Love. Contextuality test of the nonclassicality of variational quantum eigensolvers. *Phys. Rev. Lett.*, 123(20), 2019. doi:10.1103/physrevlett.123.200501.

[KN05] S. Khot and A. Naor. Nonembeddability theorems via Fourier analysis. *Math. Ann.*, 334:821–852, 2005. doi:https://doi.org/10.1007/s00208-005-0745-0.

[Koc20] L. Kocia. Improved strong simulation of universal quantum circuits. 2020. arXiv:2012.11739.

[KTP06] Z. Kominek and K. Troczka-Pawelec. Some remarks on subquadratic functions. *Dem. Math.*, 39:751–758, 2006. doi:https://doi.org/10.1515/dema-2006-0405.

[KTYI07] M. Koashi, F. Takenaga, T. Yamamoto, and N. Imoto. Quantum nonlocality without entanglement in a pair of qubits. 2007. arXiv:0709.3196.

[Lab22] F. Labib. Stabilizer rank and higher-order Fourier analysis. *Quantum*, 6:645, 2022. doi:10.22331/q-2022-02-09-645.

[Liu19] Z.-W. Liu. One-shot operational quantum resource theory. *Phys. Rev. Lett.*, 123(2), 2019. doi:10.1103/PhysRevLett.123.020401.

[LLR95] N. Linial, E. London, and Y. Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15:215–246, 1995. doi:doi.org/10.1007/BF01200757.

[LM00] N. Linial and A. Magen. Least-distortion Euclidean embeddings of graphs: products of cycles and expanders. *J. Combin. Theory Ser. B*, 79:157–171, 2000. doi:https://doi.org/10.1006/jctb.2000.1953.

[LMN02] N. Linial, A. Magen, and A. Naor. Girth and Euclidean distortion. *Geom. Funct. Anal*, 12:380–394, 2002. doi:doi.org/10.1007/s00039-002-8251-y.

[LS22] B. Lovitz and V. Steffan. New techniques for bounding stabilizer rank. *Quantum*, 6:692, 2022. doi:10.22331/q-2022-04-20-692.

[LV] M. Laurent and F. Vallentin. *A course on semidefinite optimization*. Cambridge University Press, in preparation.

[LW22] Z.-W. Liu and A. Winter. Many-body quantum magic. *PRX Quantum*, 3:020333, 2022. doi:10.1103/PRXQuantum.3.020333.

[Mat02] J. Matoušek. *Lectures on discrete geometry*. Springer, 2002. doi:doi.org/10.1007/978-1-4613-0039-7.

[ME12] A. Mari and J. Eisert. Positive Wigner functions render classical simulation of quantum computation efficient. *Phys. Rev. Lett.*, 109:230503, 2012. doi:10.1103/PhysRevLett.109.230503.

[Moo16] E. H. Moore. On properly positive Hermitian matrices. *Bull. Amer. Math. Soc.*, 23:66–67, 1916.

[Mor38] L. J. Mordell. The definite quadratic form in eight variables with determinant unity. *J. Math. Pures Appl.*, 17:41–46, 1938.

[NC11] M. A. Nielsen and I. L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, 2011. doi:10.1017/CBO9780511976667.

[Neb13] G. Nebe. Boris venkov's theory of lattices and spherical designs. *Diophantine methods, lattices, and arithmetic theory of quadratic forms, Contemp. Math., Amer. Math. Soc.*, 587:1–19, 2013. arXiv:1201.1834.

[Nie73] H. V. Niemeier. Definite quadratische formen der dimension 24 und diskriminante 1. *Journal of Number Theory*, 5:142–178, 1973. doi:doi.org/10.1016/0022-314X(73)90068-1.

[OHG] V. Obst, A. Heimendahl, and D. Gross. Wigner's theorem for stabilizer states. In preparation.

[OZR21] C. Okay, M. Zurel, and R. Raussendorf. On the extremal points of the Λ-polytopes and classical simulation of quantum computation with magic states. *Quantum Information & Computation*, 21(13&14), 2021. doi:10.26421/QIC21.13-14-2.

[Pat00] G. Pataki. The geometry of semidefinite programming. In *Handbook of semidefinite programming*, pages 29–65. Springer, 2000. doi:10.1007/978-1-4615-4381-7_3.

[PRSKB22] H. Pashayan, O. Reardon-Smith, K. Korzekwa, and S. D. Bartlett. Fast estimation of outcome probabilities for quantum circuits. *PRX Quantum*, 3:020361, 2022. doi:10.1103/PRXQuantum.3.020361.

[PSV22] S. Peleg, A. Shpilka, and B. L. Volk. Lower bounds on stabilizer rank. *Quantum*, 6:652, February 2022. doi:10.22331/q-2022-02-15-652.

[PWB15] H. Pashayan, J. J. Wallman, and S. D. Bartlett. Estimating outcome probabilities of quantum circuits using quasiprobabilities. *Phys. Rev. Lett. 115, 070501*, 2015. doi:10.1103/PhysRevLett.115.070501.

[QPG21] H. Qassim, H. Pashayan, and D. Gosset. Improved upper bounds on the stabilizer rank of magic states. *Quantum*, 5:606, 2021. doi:10.22331/q-2021-12-20-606.

[RBVT+20] R. Raussendorf, J. Bermejo-Vega, E. Tyhurst, C. Okay, and M. Zurel. Phase space simulation method for quantum computation with magic states on qubits. *Phys. Rev. A 101, 012350*, 2020. doi:10.1103/PhysRevA.101.012350.

[Reg18] B. Regula. Convex geometry of quantum resource quantification. *J. Phys. A: Math. Theor. 51, 045303*, 2018. doi:10.1088/1751-8121/aa9100.

[Rei05] B. Reichardt. Quantum universality from magic states distillation applied to css codes. *Quantum Information Processing*, 4:251–264, 2005. doi:10.1007/s11128-005-7654-8.

[RSD17] O. Regev and N. Stephens-Davidowitz. A reverse Minkowski theorem. STOC 2017, pages 941–953. Association for Computing Machinery, New York, NY, USA, 2017. doi:10.1145/3055399.3055434.

[SA04] D. Gottesman S. Aaronson. Improved simulation of stabilizer circuits. *Phys. Rev. A 70, 052328*, 2004. doi:10.1103/PhysRevA.70.052328.

[SC19] J. R. Seddon and E. T. Campbell. Quantifying magic for multi-qubit operations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 475(2227):20190251, 2019. doi:10.1098/rspa.2019.0251.

[Ser73] J. P. Serre. *A. Course in Arithmetic.* Springer, 1973. doi:doi.org/10.1007/978-1-4684-9884-4.

[Sim15] B. Simon. *Harmonic Analysis and A. Comprehensive Course in Analysis, Part 3.* Am. Math. Soc., 2015.

[SRP+21] J. R. Seddon, B. Regula, H. Pashayan, Y. Ouyang, and E. T. Campbell. Quantifying quantum speedups: Improved classical simulation from tighter magic monotones. *PRX Quantum*, 2(1):010345, 2021. doi:10.1103/PRXQuantum.2.010345.

[SS06] P. Sarnak and A. Strömbergsson. Minima of epstein's zeta function and heights of flat tori. *Inv. Math.*, 165:115–151, 2006. doi:doi.org/10.1007/s00222-005-0488-2.

[SSV12] M. Dutour Sikirić, A. Schürmann, and F. Vallentin. Inhomogeneous extreme forms. *Annales de l'institut Fourier*, 62:2227–2255, 2012. doi:10.5802/aif.2748.

[Sti] F. H. Stillinger. Phase transitions in the Gaussian core system. *J. Chem. Phys.*, pages 3968–3974. doi:https://doi.org/10.1063/1.432891.

[Tea] The Sage Development Team. *SageMath, the Sage Mathematics Software System (Version 9.1)*, volume 2020. URL: http://www.sagemath.org.

[Val08] F. Vallentin. Optimal distortion embeddings of distance regular graphs into Euclidean spaces. *J. Combin. Theory Ser. B*, 98:95–104, 2008. doi:https://doi.org/10.1016/j.jctb.2007.06.002.

[Ven80] B. B. Venkov. The classification of integral even unimodular 24-dimensional quadratic forms. *Algebra, number theory and their applications*, 148(1978):63–74, 1980.

[Ven01] B. B. Venkov. Réseaux euclidiens, designs sphériques et formes modulaires. *Monogr. Enseign. Math*, 37:10–86, 2001.

[VFGE12] V. Veitch, C. Ferrie, D. Gross, and J. Emerson. Negative quasi-probability as a resource for quantum com-

putation. *New. J. Phys.*, 14(11):113011, 2012. doi:10.1088/1367-2630/14/11/113011.

[Via17] M. Viazovska. The sphere packing problem in dimension 8. *Annals of Mathematics*, 185:991–1015, 2017. doi:https://doi.org/10.4007/annals.2017.185.3.7.

[VMGE14] V. Veitch, S. A. H. Mousavian, D. Gottesman, and J. Emerson. The resource theory of stabilizer quantum computation. *New. J. Phys.*, 16(1):013009, 2014. doi:10.1088/1367-2630/16/1/013009.

[Wid41] D. V. Widder. *The Laplace Transform*. Princeton University Press, 1941. doi:doi.org/10.1515/9781400876457-003.

[Wit41] E. Witt. Eine Identität zwischen Modulformen zweiten Grades. *Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg*, 14:323–337, 1941.

[WWS19] X. Wang, M. M. Wilde, and Y. Su. Quantifying the magic of quantum channels. *New. J. Phys.*, 21(10):103002, 2019. doi:10.1088/1367-2630/ab451d.

[WWS20] X. Wang, M. M. Wilde, and Y. Su. Efficiently computable bounds for magic state distillation. *Phys. Rev. Lett.*, 124(9):090505, 2020. doi:10.1103/PhysRevLett.124.090505.

[Zie95] G. M. Ziegler. *Lectures on Polytopes*. Springer, 1995. doi:10.1007/978-1-4613-8431-1.

[ZOR20] M. Zurel, C. Okay, and R. Raussendorf. Hidden variable model for universal quantum computation with magic states on qubits. *Phys. Rev. Lett.*, 125(26):260404, 2020. doi:10.1103/PhysRevLett.125.260404.

[ZORH21] M. Zurel, C. Okay, R. Raussendorf, and A. Heimendahl. Hidden variable model for quantum computation with

magic states on any number of qudits of any dimension. 2021. arXiv:2110.12318.

[Zur20] M. Zurel. *Hidden variable models and classical simulation algorithms for quantum computation with magic states on qubits*. Master's thesis, University of British Columbia, 2020. doi:10.14288/1.0394790.

[Zur21] M. Zurel. Private communication, 2021.

Eidesstattliche Versicherung

Gemäß der Promotionsordnung vom 12. März 2020

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel und Literatur angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind als solche kenntlich gemacht. Ich versichere an Eides statt, dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen und eingebundenen Artikeln und Manuskripten - noch nicht veröffentlicht worden ist sowie, dass ich eine Veröffentlichung der Dissertation vor Abschluss der Promotion nicht ohne Genehmigung des Promotionsausschusses vornehmen werde. Die Bestimmungen dieser Ordnung sind mir bekannt. Darüber hinaus erkläre ich hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten der Universität zu Köln gelesen und sie bei der Durchführung der Dissertation zugrundeliegenden Arbeiten und der schriftlich verfassten Dissertation beachtet habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Teilpublikationen:

1. Arne Heimendahl, Felipe Montealegre-Mora, Frank Vallentin and David Gross. “Stabilizer extent is not multiplicative”. In: *Quantum 5, 400*, 2021, `https://doi.org/10.22331/q-2021-02-24-400`
2. Arne Heimendahl, Markus Heinrich and David Gross. “The ax-

iomatic and the operational approaches to resource theories of magic do not coincide". In: Journal of Mathematical Physics 63, 112201 (2022), https://doi.org/10.1063/5.0085774

3. Arne Heimendahl, Aurelio Marafioti, Antonia Thiemeyer, Frank Vallentin and Marc Christian Zimmermann. "Critical Even Unimodular Lattices in the Gaussian Core Model." In: *International Mathematics Research Notices*, 2022, rnac164, https://doi.org/10.1093/imrn/rnac164

4. Arne Heimendahl, Moritz Lücke, Frank Vallentin, and Marc Christian Zimmermann. "A semidefinite program for least distortion embeddings of flat tori into Hilbert spaces", 2022, arXiv: 2210.11952