Divisorial gonality of graphs, the slice rank polynomial method, and tensor products of convex cones

van Dobben de Bruyn, J.

DOI
10.4233/uuid:bb2db244-e032-46bd-a9d7-a36b9ce0ce0e

Publication date
2023

Document Version
Final published version

Citation (APA)
van Dobben de Bruyn, J. (2023). *Divisorial gonality of graphs, the slice rank polynomial method, and tensor products of convex cones*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:bb2db244-e032-46bd-a9d7-a36b9ce0ce0e

Divisorial gonality of graphs, the slice rank polynomial method, and tensor products of convex cones

Divisorial gonality of graphs, the slice rank polynomial method, and tensor products of convex cones

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus, prof.dr.ir. T.H.J.J. van der Hagen
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op maandag 27 maart 2023 om 17:30 uur

door

Josse van Dobben de Bruyn

Master of Science in Mathematics, Universiteit Leiden
geboren te Leiden

Dit proefschrift is goedgekeurd door de promotoren.

Samenstelling promotiecommissie bestaat uit:

Rector magnificus	voorzitter
Prof.dr. D.C. Gijswijt	Technische Universiteit Delft, promotor
Dr.ir. O.W. van Gaans	Universiteit Leiden, copromotor

Onafhankelijke leden:

Prof.dr.ir. M.C. Veraar	Technische Universiteit Delft
Prof.dr. G.L.M. Cornelissen	Universiteit Utrecht
Dr. M.F.E. de Jeu	Universiteit Leiden
Dr. J. Briët	Centrum Wiskunde & Informatica
Dr. R.E. Morrison	Williams College, Verenigde Staten
Prof.dr.ir. K.I. Aardal	Technische Universiteit Delft, reservelid

Het onderzoek beschreven in dit proefschrift is mede gefinancierd door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO), onder projectnummer 613.009.127.

Keywords: Finite graph, metric graph, gonality, chip-firing game, treewidth, tree decomposition, monotone search strategy, slice rank, finite field, system of balanced linear equations, convex cone, partially ordered vector space, ordered tensor product, face, extremal ray, order ideal

Printed by: Ipskamp Printing

Front & Back: Drawing by F.H.J. de Paus,
Mathematical figures by J. van Dobben de Bruyn,
Design by C.J. Nogarede

ISBN 978-94-6384-425-3

An electronic version of this dissertation is available at
`https://repository.tudelft.nl/`.

Table of contents

Chapter 1

Introduction

This dissertation consists of three parts, each of which discusses a separate topic. The parts are not directly related to one another, but the overarching theme is that all parts involve aspects of combinatorics, geometry, and algebra:

- In the first part, we study divisorial gonality of graphs. This relatively new graph parameter has its roots in algebraic geometry, though we will mostly focus on combinatorial aspects.
- In the second part, we study a problem on the interface of arithmetic combinatorics and finite geometry, on avoiding affine configurations in subsets of $\mathbb{F}_q^n$. For this we use an application of the slice rank polynomial method.
- In the third part, we study tensor products of convex cones in real vector spaces. This problem mostly involves linear algebra and convex geometry (and some functional analysis), but a little bit of combinatorics does come into play when studying the face structure of the minimal/maximal cone in the tensor product.

In this chapter, we give a brief overview of the scope of each of the three parts. More detailed introductions will be given in the respective chapters.

1.1 Part I: Divisorial gonality of graphs

Since the 1980s, mathematicians have been studying a family of games known collectively as *chip-firing games*. In a chip-firing game, every node of a graph is endowed with a number of *chips*, which may be redistributed according to certain *firing rules*. The original motivation for studying these games comes from sandpile models in physics, but later variations of the game bear little resemblance of such real-world phenomena.

The simplest chip-firing game is known as the *dollar game*. In this game, the number of chips (or *dollars*) on a node is also allowed to be negative, in which case the node is said to be *in debt*. The objective of the game is to get all nodes out of debt via a sequence of *firing moves*, where the player chooses a single node and decreases the number of chips on that node by giving chips to all neighbours of that node, one chip for each edge. Whether or not it is possible to get all nodes out of debt depends on the graph and on the initial chip configuration.

Around 2007, Baker and Norine showed that the dollar game is closely related to questions in algebraic geometry. In their seminal 2007 paper [BN07], they proved a graph-theoretic analogue of the Riemann–Roch theorem from algebraic geometry, and they showed that it admits an equivalent formulation in terms of chip-firing games on graphs. This was further strengthened in another paper by Baker [Bak08], which provides a concrete way to translate between curves and graphs. These two papers marked the beginning of a period of fruitful cross-pollination between algebraic geometry, tropical geometry, and graph theory, which is still going strong today. An overview of recent developments in this field can be found in the survey [BJ16] and the expository articles [DV21b, Jen21].

The theory developed by Baker and Norine has led to many new concepts and problems in graph theory. One of these is a new graph parameter, called the (*divisorial*) *gonality*, defined as the graph-theoretic analogue of the gonality of an algebraic curve. This graph parameter has attracted attention from researchers in algebraic geometry, graph theory, and theoretical computer science, and will be the focus of Part I of this dissertation.

Our main contributions to this field are twofold. First, it was proved in 2014 that gonality is lower bounded by a well-known graph parameter called *treewidth*, but the original proof was non-constructive [DG20]. In Chapter 3, we give a constructive proof of that same fact, by providing a polynomial time algorithm that turns a positive rank divisor of degree k into a tree decomposition of width at most k.

Second, in [BN07], Baker and Norine proved combinatorial analogues of many fundamental properties of algebraic curves, such as the Riemann–Roch theorem, Clifford's theorem, and Abel–Jacobi theory. One important algebro-geometric result of which they were not able to prove a combinatorial analogue is the Brill–Noether theorem. This motivated Baker to formulate the *Brill–Noether conjecture for graphs* [Bak08, Conj. 3.9(1)], which is currently one of the main open problems in the field. In Chapter 4, we make partial progress on this problem, by showing that Baker's *subdivision conjecture*, which implies the Brill–Noether conjecture, is not true. This rules out the most obvious approach towards a proof of the Brill–Noether conjecture for graphs, and makes it unclear whether or not the latter is plausible at all.

1.2 Part II: The slice rank polynomial method

Arithmetic combinatorics is the area of mathematics that deals with questions about sizes of sets subject to certain arithmetic conditions. Well-known topics in this field include sumset estimates, where one wishes to bound the size of the sumset $A + B$ in terms of the sizes of A and B (in some discrete abelian group G), and sets without k-term arithmetic progressions. We focus on the latter.

Let G be an abelian group. A *k-term arithmetic progression in G* is a sequence of the form $P = (a, a + b, a + 2b, \ldots, a + (k - 1)b)$, for some $a, b \in G$. We say that P is *non-trivial* if $b \neq 0$, and *proper* if all entries of P are different, and we say that a subset $A \subseteq G$ *contains* P if A contains all entries of P. Further, we write $\mathbb{N}_1 = \{1, 2, 3, \ldots\}$ and $[n] = \{1, 2, \ldots, n\}$.

An important problem in arithmetic combinatorics is to bound the maximum size of a set $A \subseteq G$ which does not contain a proper k-term arithmetic progression. The first results in this direction were obtained over the integers ($G = \mathbb{Z}$), starting with van der Waerden's theorem from 1927.

Van der Waerden's Theorem ([Wae27]). *For all integers $r, k \geq 2$ there is an integer $N_{r,k}$ such that, for all $n \geq N_{r,k}$ and for every partition of $[n] = X_1 \cup \cdots \cup X_r$ into r classes, at least one of the partition classes X_i contains a proper k-term arithmetic progression.*

This was subsequently strengthened by Roth [Rot52, Rot53] (for $k = 3$) and Szemerédi [Sze69, Sze75] (for $k \geq 4$) to show that, for large enough n, every set $A \subseteq [n]$ of fixed positive density δ contains proper k-term arithmetic progressions.

Szemerédi's Theorem ([Rot52, Sze69, Sze75]). *For every $k \in \mathbb{N}_1$ and every $\delta \in (0, 1]$, there is a positive integer $N_{k,\delta}$ such that, for all $n \geq N_{k,\delta}$, every subset $A \subseteq [n]$ of size $|A| \geq \delta n$ contains a proper k-term arithmetic progression.*

Szemerédi's theorem is one of the cornerstones of arithmetic combinatorics, and is continuously being refined and extended. By now, several fundamentally different proofs of Szemerédi's theorem are known. In addition to Szemerédi's original combinatorial proof, the most notable are Furstenberg's proof using ergodic theory [Fur77, FKO82], Gowers' Fourier-analytic proof [Gow98, Gow01], and a proof using a regularity lemma for hypergraphs by Rödl, Nagle, Schacht and Skokan [NRS06, RS04, RS06] and (independently) Gowers [Gow07].

It is believed that these results can still be improved, and an important open problem is the following.

Erdős–Turán Conjecture. *Every subset $A \subseteq \mathbb{N}_1$ with $\sum_{n \in A} \frac{1}{n} = \infty$ contains arbitrarily long proper arithmetic progressions.*

A special case of this conjecture, where A is the set of all prime numbers, was settled by the celebrated Green–Tao theorem [GT08]. For general sets, a recent preprint of Bloom and Sisask [BS20] proves the conjecture for arithmetic progressions of length 3, but for longer progressions the conjecture is still wide open.

In an attempt to develop new techniques that can be used over the integers, mathematicians have also studied the similar problem where $G = \mathbb{F}_p^n$ is a vector space over the finite field $\mathbb{F}_p$ for some prime number p. In this setting, we are interested in the asymptotic behaviour as p is fixed and n goes to infinity. Here we have the following problem:

Problem (Avoiding k-term arithmetic progressions in subsets of $\mathbb{F}_p^n$). *Given a prime number p and an integer k satisfying $3 \leq k \leq p$, is there a constant $C_{p,k} < p$ such that, for all $n \in \mathbb{N}$, every set $A \subseteq \mathbb{F}_p^n$ of size at least $(C_{p,k})^n$ contains a proper k-term arithmetic progression?*

The simplest case, when $p = k = 3$, is known as the *cap set problem*. This problem has drawn considerable attention in the past, not only because it forms a finite model for problems about avoiding arithmetic progressions over the integers, but also because

it is closely related to other prominent open problems in discrete mathematics, such as the sunflower conjecture and the computational complexity of matrix multiplication [ASU13].

In a quick series of events in May 2016, the cap set problem was suddenly solved by Ellenberg and Gijswijt [EG17], building on a new technique developed earlier that month by Croot, Lev and Pach [CLP17]. Their proof uses a new variation of the *polynomial method*, a collection of techniques for solving problems in combinatorics by encoding them with polynomials.

The Ellenberg–Gijswijt proof was subsequently recast by Tao [Tao16] in terms of a new rank function for tensors, called *slice rank*, and this has since become the dominant terminology. This new set of techniques, known as the *slice rank polynomial method*, is now being applied to related problems, such as the aforementioned problems on sunflowers-free sets [NS17, Nas22] and fast matrix multiplication [BCC$^+$17]. Furthermore, it has led to the study of several other rank functions, their relation to slice rank, and their applications, including the *analytic rank* [GW11, Lov19], *partition rank* [Nas20b], *G-stable rank* [Der22], *geometric rank* [KMZ20], and *asymptotic slice rank* [Zui18, §4.6].

Although the cap set problem has been solved, the aforementioned Problem (avoiding k-term arithmetic progressions in subsets of $\mathbb{F}_p^n$) remains wide open for $k \geq 4$. This problem is believed to be beyond the reach of current (slice rank) methods.

Nonetheless, progress is being made on the broader problem of avoiding affine configurations in subsets of $\mathbb{F}_q^n$. This will be the main topic of Part II of this dissertation. Here, instead of avoiding a k-term arithmetic progression, we seek to avoid non-trivial solutions to a system of balanced linear equations, where a linear equation $b_1\boldsymbol{x_1} + \cdots + b_k\boldsymbol{x_k} = 0$ (with $b_1, \ldots, b_k \in \mathbb{F}_q$ and $\boldsymbol{x_1}, \ldots, \boldsymbol{x_k} \in \mathbb{F}_q^n$) is called *balanced* if $b_1 + \cdots + b_k = 0$. This contains the problem of avoiding k-term arithmetic progressions as a special case, because a k-term arithmetic progressions can be encoded by a balanced linear system with k variables and $k-2$ equations (for instance, the equations $\boldsymbol{x_i} - 2\boldsymbol{x_{i+1}} + \boldsymbol{x_{i+2}} = 0$ for $i \in [k-2]$).

Consider a balanced linear system

$$\begin{cases} a_{11}\boldsymbol{x_1} + \cdots + a_{1k}\boldsymbol{x_k} = 0, \\ \qquad\qquad \vdots \\ a_{m1}\boldsymbol{x_1} + \cdots + a_{mk}\boldsymbol{x_k} = 0; \end{cases} \tag{$\star$}$$

consisting of m equations in the variables $\boldsymbol{x_1}, \ldots, \boldsymbol{x_k} \in \mathbb{F}_q^n$. Note that the variables are not taken from $\mathbb{F}_q$, but from $\mathbb{F}_q^n$ as $n \to \infty$.

If $k \geq 2m+1$, then a straightforward application of the slice rank method shows that there is a constant $C_{q,m,k} < q$ such that every subset $A \subseteq \mathbb{F}_q^n$ of size $|A| \geq (C_{q,m,k})^n$ contains a solution $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}) \in A^k$ of $(\star)$ where the $\boldsymbol{x_i}$ are not all equal. This was strengthened by Mimura and Tokushige [MT19a, MT19b, MT20] and Sauermann [Sau22], who showed that, for certain specific classes of balanced linear systems, one can even find a solution $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}) \in A^k$ of $(\star)$ where the $\boldsymbol{x_i}$ are pairwise distinct.

In Chapter 6, we study similar problems. Our main contributions are twofold. First, we extend the aforementioned results of Mimura and Tokushige [MT19a, MT19b,

MT20] to a much larger class of balanced linear systems, thereby also bringing their results together under a single proof. Second, we extend this problem further to show that, for certain systems, one can even find a solution $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}) \in A^k$ of $(\star)$ that is maximally affinely independent, in the sense that the vectors $\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}$ do not satisfy any balanced linear equation that is not a linear combination of the equations in $(\star)$.

The class of linear systems that we study in Chapter 6 contains all systems studied by Mimura and Tokushige [MT19a, MT19b, MT20], but is disjoint from the class of linear systems studied by Sauermann [Sau22]. These results have since been superseded by a stronger result of Gijswijt [Gij21], which simultaneously contains the results from Chapter 6 and [Sau22] as special cases.

1.3 Part III: Tensor products of convex cones

A *convex cone* is a subset of a real vector space which contains 0 and is closed under addition and multiplication by positive scalars. Convex cones have applications in all areas of mathematics, ranging from algebraic geometry to optimization, as well as in other areas of science, ranging from quantum physics to economics.

In these applications, one frequently has to find a way to somehow carry the cone along when the containing vector space is modified. Therefore it is important to match common operations on vector spaces with appropriate operations on the convex cones contained in them. Examples of such operations are duals, projections, direct sums, and tensor products. In each of these, there is a straightforward (and canonical) way of carrying along the convex cone, except in the case of tensor products. Here there is not a single canonical candidate, but rather many "reasonable" candidates. For this reason, tensor products of convex cones are more involved, but also more interesting than most other operations on convex cones. They have been studied by many authors, and will also be the main topic of Part III of this dissertation.

Although much has already been said about tensor products of convex cones, many basic properties have so far gone unnoticed, and several basic questions remained unanswered. We aim to address these in Part III.

Our main contributions are threefold. First, our manuscript is one of the first to study the problem in full generality. Most of the existing literature either focuses on Archimedean lattice cones (in the functional analysis literature) or on closed, proper and generating cones in finite-dimensional spaces (in linear algebra and in applications in other fields). This means that many cones are not covered by either regime, including even some standard cones such as an infinite-dimensional positive semidefinite cone or a lexicographical cone. For general cones, results are few and far between, and even some basic questions remain unanswered. We address this by developing the theory of tensor products of convex cones in full generality, for arbitrary cones in arbitrary real vector spaces.

Second, apart from extending several known results to the infinite-dimensional setting, we prove many results which are altogether new. For instance, we show that the projective and injective cone satisfy mapping properties which are analogous to the mapping properties of the projective and injective *norm*, we give a direct formula for the lineality space of the projective or injective cone, we give precise necessary

and sufficient conditions for the projective or injective cone to be semisimple, and we exhibit new ways of constructing faces of the projective or injective cone from faces of the base cones. Furthermore, as an application of our results, we show that the tensor product of symmetric convex sets preserves proper faces, a result which we believe was only known for extreme points (and only under additional topological assumptions).

Third, we give many examples where the projective cone is not dense in the injective cone. This question has been studied by various authors throughout the years, and has seen a lot of progress in recent years thanks to interest from researchers in operator theory and theoretical physics (we discuss these connections in a bit more detail in §7.1). For a large class of closed, proper and generating cones in finite-dimensional spaces, we prove that the projective cone is closed and strictly contained in the injective cone, thereby confirming a conjecture of Barker for nearly all cones. However, as the manuscript upon which Part III is based was being written, our results were superseded by simultaneous work of Aubrun, Lami, Palazuelos and Plávala [ALPP21], who independently proved Barker's conjecture in full generality. We recover their result for nearly all cones, using completely different techniques.

A detailed outline of Part III will be given in Chapter 7.

Part I

Divisorial gonality of graphs

Chapter 2

Divisors, chip-firing games, and gonality

In algebraic geometry, *gonality* is an invariant that measures the complexity of an algebraic curve. In [Bak08], Matt Baker defined a combinatorial analogue of this invariant, the *divisorial gonality* of a graph, which is the main topic of Part I of this dissertation. In this introductory chapter, we cover the basics of gonality theory for graphs.

This chapter is based in part on the preliminaries of the papers [BDGS22] and [DSW22].

Introduction

In Part I, we focus on a graph parameter called *gonality*. This parameter has its origins in algebraic geometry, where the gonality of an algebraic curve is an invariant which measures the complexity of the curve. By viewing graphs as discrete analogues of algebraic curves, several authors have defined analogous notions for graphs. However, there are several different (inequivalent) notions of gonality for graphs, which stem from different (equivalent) definitions of the gonality of a curve. In this dissertation, we focus on *divisorial gonality*, which was the first notion of gonality to be defined for graphs (see [Bak08, §3]). For an overview of other notions of gonality of graphs, see for instance [CKK15, Appendix A].

In this chapter, we define all the relevant concepts behind divisorial gonality, and we prove the basic properties that we will use in the next chapters.

2.1 Graphs

Throughout this dissertation, by a *graph* we mean a finite, loopless, undirected multigraph. In other words, parallel edges are allowed, but self-loops are not. Furthermore, throughout this dissertation, we assume that all graphs are connected. The set of vertices of a graph G is denoted $V(G)$ and the set of edges is denoted $E(G)$. If there is exactly one edge between u and v, then we denote it by uv.

Let G be a graph. For (not necessarily disjoint) vertex sets $U, W \subseteq V(G)$, we denote by $E(U, W)$ the set of edges having one endpoint in U and the opposite endpoint in W.

We use the shorthand notation $E(u, W) := E(\{u\}, W)$ and $\delta(U) := E(U, V(G) \setminus U)$. By $N(U)$ we denote the set of vertices in $V(G) \setminus U$ that have a neighbour in U.

The *degree* of a vertex $v \in V(G)$ is $\deg(v) := |\delta(v)|$. Given a subset $U \subseteq V(G)$ and a vertex $v \in U$, the *out-degree of v with respect to U* is defined as $\mathrm{outdeg}_U(v) := |E(v, V(G) \setminus U)|$.

The *Laplacian* of a graph G is the matrix $L_G \in \mathbb{Q}^{V(G) \times V(G)}$ given by

$$(L_G)_{uv} = \begin{cases} \deg(u), & \text{if } u = v; \\ -|E(u, v)|, & \text{if } u \neq v. \end{cases}$$

Since we assume all graphs to be connected, the null space of L_G contains only the multiples of the all-ones vector $\mathbb{1}$.

2.2 Divisors on graphs

A *divisor* on a graph G is an element of the free abelian group on G. In other words, a divisor is a formal sum $\sum_{v \in V(G)} a_v v$, where $a_v \in \mathbb{Z}$ for all v. If D is a divisor on G and if $w \in V(G)$, then we use the notation $D(w)$ to denote the coefficient a_w of w in D. The *support* $\mathrm{supp}(D)$ of a divisor D is the set of all $v \in V(G)$ for which $D(v) \neq 0$.

For two divisors D and D', we write $D \geq D'$ if $D(v) \geq D'(v)$ for all v. A divisor D is called *effective* if $D \geq 0$. The sets of all divisors and all effective divisors on G are denoted by $\mathrm{Div}(G)$ and $\mathrm{Div}_+(G)$, respectively.

The *degree* of a divisor is the sum of its coefficients: $\deg(D) := \sum_{v \in V(G)} D(v)$. The set of all effective divisors of degree d on G is denoted $\mathrm{Div}^d_+(G)$.

The Laplacian matrix L_G of G defines a map $\mathbb{Z}^{V(G)} \to \mathrm{Div}(G)$, $x \mapsto L_G x$. Divisors in the image of this map are called *principal divisors*. Two divisors $D, D' \in \mathrm{Div}(G)$ are *equivalent*, written $D \sim D'$, if $D - D'$ is a principal divisor. This defines an equivalence relation, and the equivalence classes coincide with the cosets of the subgroup $\mathrm{Prin}(G) \subseteq \mathrm{Div}(G)$ of principal divisors. Equivalent divisors have the same degree, because $\mathbb{1}^T L_G = 0$.

Let D and D' be equivalent divisors. Then $D' = D - L_G x$ for some $x \in \mathbb{Z}^{V(G)}$, but x is not unique. Since $\ker(L_G) = \mathrm{span}(\mathbb{1})$, there is exactly one such x with the additional property that $x \geq 0$ and $x_v = 0$ for at least one $v \in V(G)$. We denote this x by $\mathrm{script}(D, D')$ and write $\mathrm{dist}(D, D') = \max\{x_v \mid v \in V(G)\}$. Note that if $t = \mathrm{dist}(D, D')$, then $\mathrm{script}(D', D) = t\mathbb{1} - x$, so $\mathrm{dist}(D', D) = \mathrm{dist}(D, D')$. Furthermore, if $D \sim D' \sim D''$, then we have the triangle inequality $\mathrm{dist}(D, D'') \leq \mathrm{dist}(D, D') + \mathrm{dist}(D', D'')$, because $\mathrm{script}(D, D'') = \mathrm{script}(D, D') + \mathrm{script}(D', D'') - c\mathbb{1}$ for some integer $c \geq 0$.

The *rank* of a divisor $D \in \mathrm{Div}(G)$ is defined as

$$\mathrm{rank}(D) := \max\{k \in \mathbb{Z} \mid D - E \text{ is equivalent to an effective divisor for all } E \in \mathrm{Div}^k_+(G)\}.$$

We have $\mathrm{rank}(D) = -1$ if and only if D is not equivalent to an effective divisor.

Given a graph G and an integer $r \geq 1$, the *r-th divisorial gonality* $\mathrm{dgon}_r(G)$ of G is the minimum degree of a rank r divisor on G. For $r = 1$, this is simply called the *divisorial gonality* of G, written $\mathrm{dgon}(G) := \mathrm{dgon}_1(G)$.

2.3 The chip-firing game

Equivalence of divisors can also be described in terms of a "chip-firing game". In this game, we interpret the divisor D as a distribution of chips over the vertices of G, where $D(v)$ is the number of chips on the vertex v. A vertex with a negative number of chips is said to be *in debt*.

In a *vertex firing move*, a vertex $v \in V(G)$ sends chips to its neighbours, one along each incident edge. This turns D into the divisor $D' = D - L_G \mathbb{1}_v$, where $\mathbb{1}_v$ is the characteristic vector of v. In a *subset firing move*, we simultaneously fire all vertices in the vertex set $U \subseteq V(G)$.[1] Edges with both endpoints in U see two chips going in opposite direction along the edge, so these cancel out. Therefore the net effect of firing U is that one chip is moved along each edge in the cut $E(U, V(G) \setminus U)$, from U to $V(G) \setminus U$. The resulting divisor is $D' = D - L_G \mathbb{1}_U$, where $\mathbb{1}_U$ denotes the characteristic vector of U. Since $L_G \mathbb{1} = 0$, firing U can be undone by firing $V(G) \setminus U$, and we call this *inverse firing* U.

When studying divisorial gonality, we often restrict our attention to effective divisors. If D is an effective divisor, we say that a subset $U \subseteq V(G)$ is *valid* (or *can be fired*) with respect to D if $D(v) \geq \mathrm{outdeg}_U(v)$ for all $v \in U$. It is easy to see that U is valid if and only if $D - L_G \mathbb{1}_U$ is effective, so that firing U does not push any vertices in debt.

The following proposition shows that it is possible to move between every pair of equivalent effective divisors by subset firing moves without ever going into debt.

Proposition 2.1 ([DG20, Lem. 2.3])**.** *Let D, D' be equivalent effective divisors. Then there is a unique increasing sequence $\varnothing \subsetneq U_1 \subseteq U_2 \subseteq \cdots \subseteq U_t \subsetneq V(G)$ of vertex sets such that subsequently firing $U_1, \ldots, U_t$ (in that order) turns D into D' without ever going into debt.*

Proof. Let $x = \mathrm{script}(D, D')$ and $t = \mathrm{dist}(D, D') = \max\{x_v \mid v \in V(G)\}$. Let $U_1 \subseteq U_2 \subseteq \cdots \subseteq U_t$ be the reverse level set decomposition of x; that is:

$$U_i := \{v \in V(G) : x_v \geq t + 1 - i\}, \qquad \text{for all } i \in [t].$$

Then $x = \sum_{i=1}^{t} \mathbb{1}_{U_i}$, so subsequently firing $U_1, \ldots, U_t$ turns D into D'. Furthermore, if $\varnothing \subsetneq U_1' \subseteq U_2' \subseteq \cdots \subseteq U_{t'}' \subsetneq V(G)$ is another increasing sequence such that subsequently firing $U_1', \ldots, U_{t'}'$ turns D into D', then $y := \sum_{i=1}^{t'} \mathbb{1}_{U_i'} \in \mathbb{Z}^{V(G)}$ satisfies $D' = D - L_G y$ as well as the additional properties $y \geq 0$ and $y_v = 0$ for at least one $v \in V(G)$ (because $U_{t'}' \subsetneq V(G)$), so we have $y = x$. But there is only one way to decompose x as the sum of characteristic vectors of an increasing sequence of vertex sets, so we have $t' = t$ and $U_i = U_i'$ for all $i \in [t]$. This proves uniqueness.

Let $D_0, \ldots, D_t$ be the sequence of intermediate divisors, so that $D_0 = D$ and $D_i = D_{i-1} - L_G \mathbb{1}_{U_i}$ for all $i \in [t]$. We must show that $D_i \geq 0$ for all $i \in [t]$. For $i = t$ this is clear, because $D_t = D'$. Now let $i \in [t-1]$ and $v \in V(G)$. To show that $D_i(v) \geq 0$, we distinguish two cases.

[1] Note that the order of firing does not matter, because addition is commutative: $D - L_G \mathbb{1}_v - L_G \mathbb{1}_u = D - L_G \mathbb{1}_u - L_G \mathbb{1}_v$.

- If $v \notin U_i$, then also $v \notin U_1, \dots, U_i$ (because $U_1 \subseteq \cdots \subseteq U_i$), so v has only received chips so far. Therefore $D_i(v) \geq D(v) \geq 0$.
- if $v \in U_i$, then also $v \in U_i, \dots, U_t$ (because $U_i \subseteq \cdots \subseteq U_t$), so v will only give away chips from now on. Hence $D_i(v) \geq D'(v) \geq 0$. ■

This leads to the following alternative definition of gonality: $\operatorname{dgon}_r(G)$ is the minimum number of chips in a chip configuration (divisor) $D \geq 0$ such that, by subsequently firing valid vertex sets, we can reach for every $E \in \operatorname{Div}_+^r(G)$ a divisor $D' \geq E$. In particular, $\operatorname{dgon}(G)$ is the minimum number of chips in a chip configuration $D \geq 0$ such that, by subsequently firing valid vertex sets, we can reach for every $v \in V(G)$ an effective divisor $D' \geq 0$ with $D'(v) \geq 1$.

2.4 Reduced divisors

Let G be a graph, and let $q \in V(G)$. An effective divisor $D \in \operatorname{Div}_+(G)$ is said to be *q-reduced* if every non-empty valid set contains q.[2] In Proposition 2.3 below we prove that every effective divisor is equivalent to exactly one q-reduced divisor. For this we use the following lemma.

Lemma 2.2. *Let $D \in \operatorname{Div}_+(G)$ be a q-reduced divisor, and let $D' \sim D$ be an effective divisor equivalent to D. Then $\operatorname{script}(D, D')_q = \operatorname{dist}(D, D')$ and $\operatorname{script}(D', D)_q = 0$.*

Proof. Write $x := \operatorname{script}(D, D')$ and $t := \operatorname{dist}(D, D')$. By the proof of Proposition 2.1, the highest level set $U_1 := \{v \in V(G) : x_v = t\}$ is non-empty and valid with respect to D. Since D is q-reduced, it follows that $q \in U_1$, so $x_q = t$. The other equality follows because $\operatorname{script}(D, D') + \operatorname{script}(D', D) = t\mathbb{1}$. ■

Proposition 2.3 ([BN07, Prop. 3.1])**.** *For every $D \in \operatorname{Div}_+(G)$, there is a unique q-reduced divisor $D' \in \operatorname{Div}_+(G)$ such that $D \sim D'$.*

Proof. To prove existence, we construct a sequence $D_0, D_1, \dots, D_n, \dots \in \operatorname{Div}_+(G)$ of effective divisors equivalent to D recursively in the following way:

- Set $D_0 := D$.
- Suppose that $D_0, \dots, D_n$ have been defined. If D_n is q-reduced, set $D' := D_n$ and terminate. Otherwise, choose a non-empty subset $U_n \subseteq V(G) \setminus \{q\}$ that is valid with respect to D_n and set $D_{n+1} := D_n - L_G \mathbb{1}_{U_n}$. Repeat.

Since D_{n+1} is obtained from D_n by firing a valid vertex set, we have $D_i \geq 0$ for all i, and $D_0 \sim D_1 \sim \cdots \sim D_n$. Furthermore, since $q \notin U_0, \dots, U_{n-1}$, we have $\operatorname{script}(D_0, D_n) = \sum_{i=0}^{n-1} \mathbb{1}_{U_i}$. In particular, for $n \neq m$ we have $\operatorname{script}(D_0, D_n) \neq \operatorname{script}(D_0, D_m)$, and therefore $D_n \neq D_m$. Since there are only finitely many effective divisors of degree

[2] There is also a notion of q-reduced divisors which are not effective; see [BN07, §3.1]. For simplicity, we restrict our attention to effective divisors, which suffice for our purposes. Here the definitions and proofs are slightly simpler. It is not hard to see that our notion of q-reduced divisors agress with the one from [BN07, §3.1].

$\deg(D)$, the algorithm must terminate at some point, which shows that D is equivalent to a q-reduced divisor.

To prove uniqueness, suppose that $D', D'' \in \mathrm{Div}_+(G)$ are q-reduced effective divisors equivalent to D. Since both D' and D'' are q-reduced, it follows from Lemma 2.2 that $\mathrm{script}(D', D'')_q = \mathrm{dist}(D', D'')$, but also $\mathrm{script}(D', D'')_q = 0$. Therefore $\mathrm{dist}(D', D'') = 0$, hence $D' = D''$. ■

The unique q-reduced divisor equivalent to D is denoted D_q. The following proposition shows that D_q maximizes the value of $D'(q)$ among all effective divisors $D' \sim D$.

Proposition 2.4. *Let $D \in \mathrm{Div}_+(G)$ be an effective divisor, and let $q \in V(G)$. Then for every effective divisor $D' \sim D$, one has $D_q(q) \geq D'(q)$.*

Proof. Let $D' \sim D$ be effective, and write $x := \mathrm{script}(D', D_q)$. Then $x \geq 0$ and $x_q = 0$, by Lemma 2.2. Hence, when moving from D' to D_q by subsequently firing the level sets $U_1, \ldots, U_t$ (see Proposition 2.1), the vertex q only receives chips, never giving anything away. Therefore $D_q(q) \geq D'(q)$. ■

Corollary 2.5. *Let $D \in \mathrm{Div}_+(G)$. Then $\mathrm{rank}(D) \geq 1$ if and only if $D_q(q) \geq 1$ for all $q \in V(G)$.*

The following estimate of the distance between D and D_q will be useful later on.

Proposition 2.6. *Let $D \in \mathrm{Div}_+(G)$ be an effective divisor and let $q \in V(G)$. Then $\mathrm{dist}(D, D_q) \leq \deg(D) \cdot |V(G)|$.*

Proof. By Proposition 2.1, there is an increasing sequence $\varnothing \subsetneq U_1 \subseteq U_2 \subseteq \cdots \subseteq U_t \subsetneq V(G)$ of vertex sets such that subsequently firing $U_1, \ldots, U_t$ (in that order) turns D into D_q without ever going into debt, where $t := \mathrm{dist}(D, D_q)$. In this sequence, the same set $U_i = U_{i+1} = \cdots$ can occur at most $\deg(D)$ times in a row, because every time we fire U_i at least one chip leaves U_i. It follows that $t \leq \deg(D) \cdot |V(G)|$. ■

2.5 Dhar's burning algorithm

Dhar's burning algorithm [Dha90], given in Algorithm 2.7 below, takes as input a graph G, a divisor D and a vertex q, and returns a valid vertex set $U \subseteq V(G) \setminus \{q\}$.

Input : A triple (G, D, q), where G is a graph, $D \in \mathrm{Div}_+(G)$, and $q \in V(G)$.
Output : The maximal valid subset $U \subseteq V(G) \setminus \{q\}$.

```
1 Function DHAR(G, D, q):
2 |   U := V(G) \ {q};
3 |   while D(v) < outdeg_U(v) for some v ∈ U do
4 |   |   U := U \ {v};
5 |   end while
6 |   return U
```

Algorithm 2.7: Dhar's burning algorithm for finite graphs.

Dhar's burning algorithm can be implemented efficiently as a breadth first search. By remembering the values of $\mathrm{outdeg}_U(v)$ for all $v \in U$ (instead of computing them when needed) and only updating those values that are changed when some vertex v is removed from U, the algorithm can be implemented to run in $O(|E(G)|)$ time. For details, see [Dob12, Alg. 5.3].

We proceed to prove the basic properties of Dhar's burning algorithm.

Proposition 2.8. *The vertex set $U \subseteq V(G)\setminus\{q\}$ returned by Dhar's burning algorithm is valid and contains every other valid vertex set $U' \subseteq V(G) \setminus \{q\}$.*

Proof. When Dhar's burning algorithm terminates, it returns a set $U \subseteq V(G) \setminus \{q\}$ such that $D(v) \geq \mathrm{outdeg}_U(v)$ for all $v \in U$, so this set is (by definition) valid.

Let $U' \subseteq V(G) \setminus \{q\}$ be another valid vertex set. At the start of the algorithm, we have $U := V(G) \setminus \{q\}$, so at this point $U' \subseteq U$. Furthermore, as long as the inclusion $U' \subseteq U$ is maintained, we have $D(v) \geq \mathrm{outdeg}_{U'}(v) \geq \mathrm{outdeg}_U(v)$ for all $v \in U'$. Therefore the algorithm never removes a vertex $v \in U'$ from U. ■

In other words, Dhar's burning algorithm always returns the unique inclusionwise maximal valid set $U \subseteq V(G) \setminus \{q\}$. The following corollaries are immediate.

Corollary 2.9. *The output of Dhar's burning algorithm does not depend on the order in which vertices are selected in the while loop in lines 3–5.*

Corollary 2.10. *Let $D \in \mathrm{Div}_+(G)$ be an effective divisor, and let $q \in V(G)$. Then D is q-reduced if and only if $\textsc{Dhar}(G, D, q) = \varnothing$.*

If D is not q-reduced, then Dhar's burning algorithm can be used to reduce the distance between D and D_q.

Proposition 2.11. *Let $D \in \mathrm{Div}_+(G)$ be an effective divisor, and let $q \in V(G)$. Moreover, let $U := \textsc{Dhar}(G, D, q)$, and suppose that $U \neq \varnothing$. Then $\mathrm{dist}(D - L_G\mathbb{1}_U, D_q) = \mathrm{dist}(D, D_q) - 1$.*

Proof. Write $x := \mathrm{script}(D, D_q)$ and $t := \mathrm{dist}(D, D_q)$. By the proof of Proposition 2.1, the highest level set $U_1 := \{v \in V(G) : x_v = t\}$ is valid with respect to D, so it follows from Proposition 2.8 that $U_1 \subseteq U$.

Let $D' := D - L_G\mathbb{1}_U$ (which is effective by Proposition 2.8), and write $x' := \mathrm{script}(D', D_q)$. Then $x - \mathbb{1}_U$ and x' differ by a multiple of $\mathbb{1}$. However, it follows from Lemma 2.2 that $x_q = x'_q = 0$, and we have $(\mathbb{1}_U)_q = 0$, so in fact $x' = x - \mathbb{1}_U$. Since $U_1 \subseteq U$, it is clear that $\mathrm{dist}(D', D_q) = \max\{x'_v : v \in V(G)\} = t - 1$. ■

By Proposition 2.6 and Proposition 2.11, we need at most $\deg(D)\cdot|V(G)|$ iterations of Dhar's burning algorithm to find the unique q-reduced divisor $D_q \sim D$. Dhar's burning algorithm and a single subset firing move can both be done in $O(|E(G)|)$ time, so computing D_q from D can be done in $O(\deg(D) \cdot |V(G)| \cdot |E(G)|)$ time.

Chapter 3

Constructing tree decompositions of graphs with bounded gonality

In 2014, Gijswijt and the author showed that treewidth is a lower bound for graph gonality. In this chapter, we give a constructive proof of the same fact, by giving a polynomial-time algorithm that turns a positive rank divisor of degree d into a tree decomposition of width at most d.

This chapter is based on the paper [BDGS22]. A preliminary version of this paper appeared earlier as a conference paper in [BDGS20]. This is joint work with Hans L. Bodlaender, Dion Gijswijt, and Harry Smit.

3.1 Introduction

In the paper [DG20], originally written in 2014, van Dobben de Bruyn and Gijswijt showed that gonality is closely related to *treewidth*, a graph parameter that plays an important role in structural graph theory and theoretical computer science (See §3.2 for a definition of treewidth.) We proved that $\operatorname{dgon}(G) \geq \operatorname{tw}(G)$ for every graph G, and we gave multiple examples where the two are actually equal. However, in general the two can be arbitrarily far apart, as can be seen by taking a "chain of cycles"; see [CDPR12]. (For another construction, see [Hen18].)

Given their very different origins, it is rather surprising that gonality and treewidth are so closely related. Our original proof from [DG20] does little to clarify this connection, as it is non-constructive and makes use of a dual characterization of treewidth in terms of brambles. In this chapter, we intend to clarify this connection by providing a constructive proof of the same fact. Specifically, we give a polynomial-time algorithm that turns a positive rank divisor of degree k into a tree decomposition of width at most k. Our main result is the following.

Theorem 3.1. *There is an $O(k \cdot |V(G)|^2 \cdot |E(G)|)$ time algorithm that takes as input a graph G and a positive rank effective divisor of degree k defined on G, and returns as output a tree decomposition of G of width at most k.*

To prove Theorem 3.1, we make use of an equivalent definition of treewidth in terms of a Cops and Robbers game. The main idea behind the algorithm is to use the chip-firing game to guide the searchers through the graph and capture the fugitive.

This leads to an algorithm that converts a positive rank effective divisor of degree k to a monotone search strategy with $k+1$ searchers (see §3.4). By encoding this monotone search strategy in a specific way (see §3.3), it can easily be turned into a tree decomposition of width at most k (see §3.5).

3.2 Treewidth

Before diving into the proof of Theorem 3.1, we briefly recall the definition of treewidth.

Treewidth is a graph parameter with a long history. Its first appearance was under the name of *dimension*, in 1972, by Bertelè and Brioschi [BB72]. It was rediscovered several times since, each time under a different name. (For an overview, see for instance [Bod98].)

The terms "tree decomposition" and "treewidth" were introduced by Robertson and Seymour [RS86a] as part of their fundamental work on graph minors. These notions have since become the dominant terminology. The treewidth of a graph is defined as follows.

Definition 3.2. Let G be a graph. A *tree decomposition* of G is a tuple $(T, (X_t)_{t\in V(T)})$, where T is a tree and $(X_t)_{t\in V(T)}$ is a collection of vertex sets $X_t \subseteq V(G)$, one for each node of T, that satisfies the following conditions:

(a) $\bigcup_{t\in V(T)} X_t = V$;

(b) for every edge $e \in E(G)$ with endpoints $u, v \in V(G)$, there is some $t \in V(T)$ such that $u, v \in X_t$;

(c) for every $v \in V(G)$, the set of nodes $T_v = \{t \in V(T) : v \in X_t\}$ is connected (in other words, it induces a subtree of T).

The sets $X_t \subseteq V(G)$ are called the *bags* of the tree decomposition. The *width* of the tree decomposition is $\max_{t\in V(T)} |X_t| - 1$.

The *treewidth* of G, denoted $\mathrm{tw}(G)$, is the minimum width of a tree decomposition of G.

There are several alternative (equivalent) definitions of treewidth. We will use a notion that is based on a Cops and Robbers game, introduced by Seymour and Thomas [ST93]. Here, a number of searchers need to catch a fugitive, subject to certain rules.

In the Cops and Robbers game for treewidth, searchers can move from a vertex in the graph to a "helicopter", or from a helicopter to any vertex in the graph. Between moves of searchers, the fugitive can move with infinite speed in the graph, but may not move over or to vertices with a searcher. The fugitive is captured when a searcher moves to the vertex with the fugitive, and there is no other vertex without a searcher that the fugitive can move to. The location of the fugitive is known to the searchers at all times.

We say that k *searchers can capture a fugitive in* G if there is a strategy for k searchers on G that guarantees that the fugitive is captured. In the initial configuration, the fugitive can choose a vertex, and all searchers are in a helicopter. A search strategy

is *monotone* if it is never possible for the fugitive to move to a vertex that had been unreachable before. In particular, in a monotone search strategy, there is never a path without searchers from the location of the fugitive to a vertex previously occupied by a searcher.

Theorem 3.3 (Seymour and Thomas [ST93])**.** *Let G be a graph and k a positive integer. The following statements are equivalent.*

(i) The treewidth of G is at most k.

(ii) $k+1$ searchers can capture a fugitive in G.

(iii) $k+1$ searchers can capture a fugitive in G with a monotone search strategy.

3.3 Monotone search strategies

We start by providing a way to encode monotone search strategies. Let G be a graph. For $X \subseteq V(G)$, the vertex set of a component of $G \setminus X$ is called an *X-flap*. A *position* is a pair (X, R), where $X \subseteq V(G)$ and R is a union[1] of X-flaps (we allow $R = \varnothing$). Note that R is a union of X-flaps if and only if $N(R) \subseteq X$.

The set X represents the vertices occupied by searchers, and the fugitive can move freely within some X-flap contained in R (if $R = \varnothing$, then the fugitive has been captured). In a monotone search strategy, the fugitive will remain confined to R, so placing searchers on vertices other than R is of no use. Therefore, it suffices to consider three types of moves for the searchers: (a) remove searchers that are not necessary to confine the fugitive to R; (b) add searchers to R; (c) if R consists of more than one X-flap, restrict attention to the X-flap $R_i \subsetneq R$ containing the fugitive. This leads us to the following definition.

Definition 3.4. Let G be a graph and let k be a positive integer. A *monotone search strategy* (MSS) with k searchers for G is a directed tree $T = (\mathcal{P}, F)$ where $\mathcal{P}$ is a set of positions with $|X| \leq k$ for every $(X, R) \in \mathcal{P}$, that satisfies the following additional conditions:

(i) The root of T is $(\varnothing, V)$.

(ii) If (X, R) is a leaf of T, then $R = \varnothing$.

(iii) Let (X, R) be a non-leaf of T. Then $R \neq \varnothing$ and there is a set $X' \subseteq X \cup R$ such that exactly one of the following applies:

  (a) $X' \subsetneq X$, and the position (X', R) is the unique out-neighbour of (X, R);

  (b) $X' \supsetneq X$, and the position $(X', R \setminus X')$ is the unique out-neighbour of (X, R);

  (c) $X' = X$, and the out-neighbours of (X, R) are the positions $(X, R_1), \ldots, (X, R_t)$ where $t \geq 2$ and $R_1, \ldots, R_t$ are the X-flaps contained in R.

[1] Here we deviate from the definition of position as stated in [ST93], in that we allow R to consist of zero X-flaps or more than one X-flap.

If condition (ii) does not necessarily hold, we say that T is a *partial* MSS. Note that we do not consider the root node to be a leaf even if it has degree 1.

It is clear that if T is an MSS for k searchers, then, as the name suggests, k searchers can capture the fugitive, the fugitive can never reach a vertex that it could not reach before, and a searcher is never placed on a vertex from which a searcher was previously removed.

We should point out that Definition 3.4 is slightly different from existing (formal or informal) definitions of a monotone search strategy in the literature. Compared to Seymour and Thomas [ST93], we also allow positions that consist of zero X-flaps or more than one X-flap. We do not prove that Theorem 3.3 also holds for our definition of a monotone search strategy (Definition 3.4), but we will show in §3.5 that an MSS with k searchers yields a tree decomposition of width at most $k-1$ in a relatively straightforward fashion.

First we focus on constructing an MSS in polynomial time. For this we use the following lemmas.

Lemma 3.5. *Let G be a graph and let T be a partial MSS with k searchers for G. Then T has at most $|V(G)|^2+1$ nodes.*

Proof. For every position (X,R), define $f(X,R)=|R|(|X|+|R|)$. We claim that for every non-leaf node (X,R), the value of $f(X,R)$ is at least the sum of the values of its children plus the number of children. In case (a) and (b), the node (X,R) has exactly one child (X',R'), which satisfies $R'\subseteq R$ and $X'\cup R'\subseteq X\cup R$, with at least one of these inclusions strict (and $|R|\neq 0$). Therefore we have $f(X,R)>f(X',R')$, hence $f(X,R)\geq f(X',R')+1$. Moreover, in case (c), we have $f(X,R)\geq f(X,R_1)+\cdots+f(X,R_t)+t$, because

$$\begin{aligned} f(X,R)-(f(X,R_1)+\cdots+f(X,R_t)) &= \sum_{1\leq i<j\leq t} 2|R_i|\cdot|R_j| \\ &\geq 2|R_1|\cdot(|R_2|+\cdots+|R_t|) \\ &\geq 2(t-1)\geq t. \end{aligned}$$

This proves our claim. It follows by induction that $f(X,R)$ is an upper bound on the number of descendants of (X,R) in T. Since every non-root node is a descendant of the root, it follows that the total number of nodes is at most $1+f(\varnothing,V)=1+|V(G)|^2$. ■

3.4 Construction of a monotone search strategy

In this section, we present an algorithm that turns a positive rank effective divisor of degree k into a monotone search strategy (MSS) with $k+1$ searchers. This will be the main component in our proof of Theorem 3.1. The procedure to convert an MSS into a tree decomposition is postponed until the next section.

The main idea behind our algorithm is to guide the searchers based on the way the chips move through the graph in the chip-firing game. The input divisor D provides the initial position for the searchers, after which we repeatedly use Dhar's burning

algorithm (see Algorithm 2.7) to compute the next move. By doing so in a careful way, we can compute an MSS in polynomial time. We will make this idea precise in Theorem 3.8 below. For this, we need the following lemmas.

Lemma 3.6. *Let G be a graph, let $X \subseteq V(G)$ be a non-empty vertex set, and let R be an X-flap. If $D \in \mathrm{Div}(G)$ is a positive rank effective divisor with $\mathrm{supp}(D) \cap R = \varnothing$, then the function calls $\textsc{Dhar}(G, D, q)$ and $\textsc{Dhar}(G, D, q')$ return the same non-empty set for all $q, q' \in R$.*

Proof. Let $q, q' \in R$ be arbitrary, and let U, U' be the sets returned by $\textsc{Dhar}(G, D, q)$ and $\textsc{Dhar}(G, D, q')$, respectively. Since D has positive rank and $D(q) = 0$, the divisor D is not q-reduced, so $U \neq \varnothing$. Moreover, since R is an X-flap, there is a path from q to q' in R. Hence, in Dhar's burning algorithm, if q is burned, then also q' is burned, because $\mathrm{supp}(D) \cap R = \varnothing$. Therefore $q' \notin U$. Since U' is the maximal subset $S \subseteq V(G) \setminus \{q'\}$ that can be fired, we have $U \subseteq U'$. By symmetry, we also have $U' \subseteq U$, which shows that $U = U'$. ■

By a slight abuse of notation, if D satisfies the condition from Lemma 3.6, then we denote the set returned by $\textsc{Dhar}(G, D, q)$ for any $q \in R$ as $\textsc{Dhar}(G, D, R)$, and we call this the *R-Dhar set for D*.

Lemma 3.7. *Let G be a graph, let $X \subseteq V(G)$ be a non-empty vertex set, and let R be an X-flap. If $D \in \mathrm{Div}(G)$ is a positive rank effective divisor such that $\mathrm{supp}(D) \cap (X \cup R) = X$, then by repeatedly firing R-Dhar sets, we obtain in a finite number of steps an effective divisor $D' \sim D$ such that $\mathrm{supp}(D') \cap (X \cup R) = X$ and such that the R-Dhar set $U := \textsc{Dhar}(G, D', R)$ satisfies $U \cap R = \varnothing$ and $U \cap X \neq \varnothing$.*

Proof. By Lemma 3.6, the R-Dhar set $U := \textsc{Dhar}(G, D, R)$ is non-empty. For every $q \in R$ we have $U = \textsc{Dhar}(G, D, q)$, and therefore $q \notin U$, so $U \cap R = \varnothing$. If $U \cap X \neq \varnothing$, we set $D' := D$ and we are done. Otherwise, we replace D by $D - L_G \mathbb{1}_U$ and iterate. Since $U \cap X = \varnothing$, the vertices in X do not give away chips. Moreover, since R is an X-flap, we have $N(R) \subseteq X$, so the vertices in R do not receive any chips. Therefore the property that $\mathrm{supp}(D) \cap (X \cup R) = X$ is maintained. By Proposition 2.6 and Proposition 2.11, we finish in no more than $k \cdot |V(G)|$ iterations. ■

Theorem 3.8. *There is an $O(k \cdot |V(G)|^2 \cdot |E(G)|)$ time algorithm that takes as input a graph G and a positive rank effective divisor of degree k defined on G, and returns as output an MSS with $k + 1$ searchers for G.*

Proof. Throughout the execution of the algorithm, we will keep a partial MSS, which we extend until it is an MSS. We start with only two nodes, namely the root $(\varnothing, V)$ and its child $(\mathrm{supp}(D), V \setminus \mathrm{supp}(D))$, after which we repeatedly process all leaves (X, R) with $R \neq \varnothing$ until no such nodes are left. At each step, for every leaf (X, R) of T we keep an effective divisor $D' \sim D$ such that $\mathrm{supp}(D') \cap (X \cup R) = X$ (equivalently: $X \subseteq \mathrm{supp}(D')$ and $R \cap \mathrm{supp}(D') = \varnothing$).

We now describe the iterative procedure. While T has a leaf (X, R) with $R \neq \varnothing$, let D' be the divisor associated to (X, R) and perform one of the following steps.

I. If R consists of multiple X-flaps $R_1, \dots, R_t$, then we add the nodes $(X, R_1), \dots, (X, R_t)$ as children of (X, R), associate D' to each of these nodes, and iterate.

II. If $X' := N(R)$ is a strict subset of X, then add the node (X', R) as a child of (X, R), associate D' to this node, and iterate.

III. The remaining case is that $N(R) = X$ and R is a single X-flap. By Lemma 3.7, by repeatedly firing R-Dhar sets, we may obtain an effective divisor $D'' \sim D'$ such that $\operatorname{supp}(D'') \cap (X \cup R) = X$ and from D'' we can fire a set U such that $U \cap R = \varnothing$ and $U \cap X \neq \varnothing$. To compute the next step in our search strategy, we fire the set U, and use the chips that are fired from X to R to guide the searchers towards the fugitive. We now make this idea precise.

Write $U \cap X = \{s_1, s_2, \dots, s_t\}$, where $t = |U \cap X|$. Since we can fire on U, we have

$$D''(s_i) \geq \operatorname{outdeg}_U(s_i) \geq |N(s_i) \cap R| \qquad \text{for all } i \in \{1, \dots, t\}. \tag{3.9}$$

We write $(X'_0, R_0) := (X, R)$, and we define the positions $(X_1, R_1), (X'_1, R_1), \dots, (X_t, R_t), (X'_t, R_t)$ recursively as follows:

$$\begin{aligned} X_i &:= X'_{i-1} \cup (N(s_i) \cap R); \\ X'_i &:= X_i \setminus \{s_i\}; \\ R_i &:= R \setminus X_i. \end{aligned}$$

It is not hard to see that every edge in $\delta(R)$ has at least one endpoint in every X'_i, and that therefore $N(R_i) \subseteq X'_i \subseteq X_i$ for all i. This shows that every R_i is a union of X'_i-flaps and of X_i-flaps, so $(X_1, R_1), (X'_1, R_1), \dots, (X_t, R_t), (X'_t, R_t)$ are valid positions.

We add the path $(X, R) \to (X_1, R_1) \to (X'_1, R_1) \to \cdots \to (X_t, R_t) \to (X'_t, R_t)$ to T. It may happen that $(X_i, R_i) = (X'_{i-1}, R_{i-1})$ for some i, in which case we remove one of the two. After that, it is easy to see that the added path is valid, as every non-leaf node in the path satisfies the condition from Definition 3.4(iii)(a) or (b).

Next, we show that the added path uses at most $k + 1$ searchers. Using (3.9) and the fact that $X \subseteq \operatorname{supp}(D'')$, we see that

$$\begin{aligned} |X'_i| &= |X \setminus \{s_1, \dots, s_i\}| + |(N(s_1) \cap R) \cup \cdots \cup (N(s_i) \cap R)| \\ &\leq |X \setminus \{s_1, \dots, s_i\}| + |N(s_1) \cap R| + \cdots + |N(s_i) \cap R)| \\ &\leq \sum_{v \in X \setminus \{s_1, \dots, s_i\}} D''(v) + \sum_{v \in \{s_1, \dots, s_i\}} D''(v) \\ &= \sum_{v \in X} D''(v) \end{aligned}$$

for all i. Therefore we have $|X'_i| \leq k$ and $|X_i| \leq k + 1$ for all $i \in \{1, \dots, t\}$, so the number of searchers is at most $k + 1$.

To the leaf (X'_t, R_t) we associate the divisor $D'' - L_G\mathbb{1}_U$ (which is effective, because U can be fired). We prove that $\operatorname{supp}(D'' - L_G\mathbb{1}_U) \cap (X'_t \cup R_t) = X'_t$. To that end, let $v \in X'_t \cup R_t$. We distinguish three cases.

- If $v \in (N(s_1) \cap R) \cup \cdots \cup (N(s_t) \cap R)$, then $v \notin U$ but v has a neighbour in U, so v receives at least one chip. Therefore $v \in \operatorname{supp}(D'' - L_G\mathbb{1}_U)$.
- If $v \in X \setminus \{s_1, \ldots, s_t\}$, then $v \notin U$ but $v \in X \subseteq \operatorname{supp}(D'')$, so v starts with a chip and does not give anything away. Therefore $v \in \operatorname{supp}(D'' - L_G\mathbb{1}_U)$.
- If $v \in R_t$, then v had no chips to begin with, and v does not have a neighbour in U. Therefore $v \notin \operatorname{supp}(D'' - L_G\mathbb{1}_U)$.

Since $X'_t = (X \setminus \{s_1, \ldots, s_t\}) \cup (N(s_1) \cap R) \cup \cdots \cup (N(s_t) \cap R)$ and $X'_t \cap R_t = \varnothing$, this shows that $\operatorname{supp}(D'' - L_G\mathbb{1}_U) \cap (X'_t \cup R_t) = X'_t$.

By Lemma 3.5, the iteration terminates after at most $|V(G)|^2 + 1$ steps. This completes the construction of the monotone search strategy.

To complete the proof, we show that this algorithm runs in $O(k \cdot |V(G)|^2 \cdot |E(G)|)$ time. Since the MSS has at most $|V(G)|^2 + 1$ nodes, and each node carries $O(|V(G)|)$ data that needs to be updated, the algorithm spends $O(|V(G)|^3)$ time maintaining the MSS. This is well within the desired $O(k \cdot |V(G)|^2 \cdot |E(G)|)$ bound.

It remains to bound the number of calls to Dhar's burning algorithm. Let $\mathcal{S} \subseteq V(T)$ denote the set of nodes for which step III of the algorithm is entered (i.e. for which $N(R) = X$ and R is a single X-flap). Let $\mathcal{S}' \subseteq \mathcal{S}$ denote the set of nodes in $\mathcal{S}$ which do not have a descendant in $\mathcal{S}$, and write $\mathcal{S}' = \{(X_1, R_1), \ldots, (X_t, R_t)\}$, where $t = |\mathcal{S}'|$. For all $i \neq j$ we have $R_i \cap R_j = \varnothing$, because the children of a branch vertex of T have distinct R-sets (by Definition 3.4(iii)(c)). Therefore, $t \leq |V(G)|$.

For $i \in [t]$, let $\mathcal{S}_i$ denote the set of all nodes in $\mathcal{S}$ of which (X_i, R_i) is a descendant, including the node (X_i, R_i) itself. Furthermore, choose some $q_i \in R_i$, and let D_{q_i} denote the unique q_i-reduced divisor equivalent to D. By monotonicity of the search strategy (see Definition 3.4(iii)), we have $R_i \subseteq R$ for all $(X, R) \in \mathcal{S}_i$, so in particular $q_i \in R$ for all $(X, R) \in \mathcal{S}_i$. Hence, by Proposition 2.11, every call to Dhar's burning algorithm made during the processing of the nodes in $\mathcal{S}_i$ decreases the distance between D and D_{q_i} by one. Since no other operations are performed on the intermediate divisors, and since $\operatorname{dist}(D, D_{q_i}) \leq k \cdot |V(G)|$ (by Proposition 2.6), it follows that the nodes in $\mathcal{S}_i$ account for at most $k \cdot |V(G)|$ calls to Dhar's burning algorithm. Therefore the total number of calls to Dhar's burning algorithm is at most $kt \cdot |V(G)| \leq k \cdot |V(G)|^2$. Dhar's burning algorithm runs in time $O(|E(G)|)$, so our algorithm spends $O(k \cdot |V(G)|^2 \cdot |E(G)|)$ time on calls to Dhar's burning algorithm. ■

3.5 Construction of a tree decomposition

To complete the algorithm, we need to turn the MSS into a tree decomposition. This is easy, as the following lemma shows.

Proposition 3.10. *Let G be a graph, and let $T' = (\mathcal{P}, F)$ be a monotone search strategy for k searchers in G. If T is the undirected tree obtained by ignoring the*

orientation of edges in T', then $(T, (X)_{(X,R)\in\mathcal{P}})$ is a tree decomposition of G of width at most $k-1$.

Proof. Let $v \in V$. We first show that $v \in X$ for some $(X, R) \in \mathcal{P}$. Let $\mathcal{P}' := \{(X, R) \in \mathcal{P} : v \in R\}$. Note that $\mathcal{P}'$ contains the root node $(\varnothing, V)$. Let $(X, R) \in \mathcal{P}'$ have maximum distance from the root. Since $v \in R$, it follows from the definition of MSS that (X, R) has a child (X', R') with $v \in X' \cup R'$. Hence, by the maximality assumption, we have $v \in X'$.

Next, we show that the set of nodes $\{(X, R) \in \mathcal{P} : v \in X\}$ is a subtree of T. Equivalently, we must show that if node (X_2, R_2) lies on a path from (X_1, R_1) to (X_3, R_3) in T, then $X_1 \cap X_3 \subseteq X_2$. It suffices to check this in two cases: the case that (X_3, R_3) is a descendant of (X_1, R_1) in T', and the case that (X_2, R_2) is the last common ancestor of (X_1, R_1) and (X_3, R_3). In the first case, it is easy to see that $X_3 \subset X_2 \cup R_2$ and $R_2 \subseteq R_1$. It follows that

$$X_1 \cap X_3 \subseteq X_1 \cap (X_2 \cup R_2) \subseteq X_1 \cap (X_2 \cup R_1) \subseteq X_2$$

since X_1 and R_1 are disjoint. In the second case, node (X_2, R_2) has more than one out-neighbour, so its out-neighbours are positions (X_2, R), where R runs over the X_2-flaps contained in R_2. It follows that $X_1 \subseteq X_2 \cup R'$ and $X_3 \subseteq X_2 \cup R''$ for distinct X_2-flaps R' and R''. Hence, $X_1 \cap X_3 \subset X_2$.

To complete the proof, it suffices to show that for every edge $e \in E(G)$ there is some node (X, R) of T such that X contains both endpoints of e. Suppose for contradiction that this is not the case for some edge e with endpoints u and v.

We first show that there is a node (X, R) such that $u \in X$ and $v \in R$ (or vice versa). To this end, consider the nodes (X, R) of T with $u, v \in R$ (e.g. the root node), and take such a node that has maximum distance from the root. This node cannot be a leaf since R is non-empty. Since u and v are neighbours, they belong to the same X-flap, so it follows by the maximum distance assumption that (X, R) has a child (X', R') with $u \in X'$ and $v \in R'$ (or vice versa).

Now consider all nodes (X, R) with $u \in X$ and $v \in R$ and take such a node for which the distance to the root is maximal. This node cannot be a leaf, because R is non-empty. Consider a child (X', R') of (X, R). If we are in case (iii)(a) then $v \in R'$, and we must have $u \in X'$, for otherwise R' is not a union of X'-flaps (since u and v are neighbours, but $v \in R'$ and $u \notin X' \cup R'$). This contradicts the maximum distance assumption. If we are in case (iii)(b), then $u \in X'$ and $v \in R'$, which once again contradicts the maximum distance assumption. If we are in case (iii)(c), we may assume that R' is the X-flap containing v and again this contradicts the maximum distance assumption. ■

Theorem 3.1 now follows from Theorem 3.8 and Proposition 3.10.

3.6 A worked example

We apply the constructions of the previous section to a relatively small example. Let $G = (V, E)$ be the graph depicted in Figure 3.1 and let $D = 3a$ be the divisor on G that has value 3 on vertex a and value 0 elsewhere.

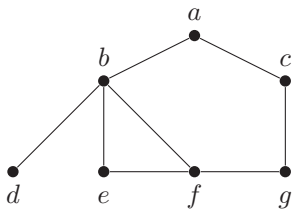


Figure 3.1: An example graph G. The divisor $D = 3a$ has positive rank. (It is not optimal, as $\mathrm{dgon}(G) = 2$, since the divisor $b + f$ also has positive rank.)

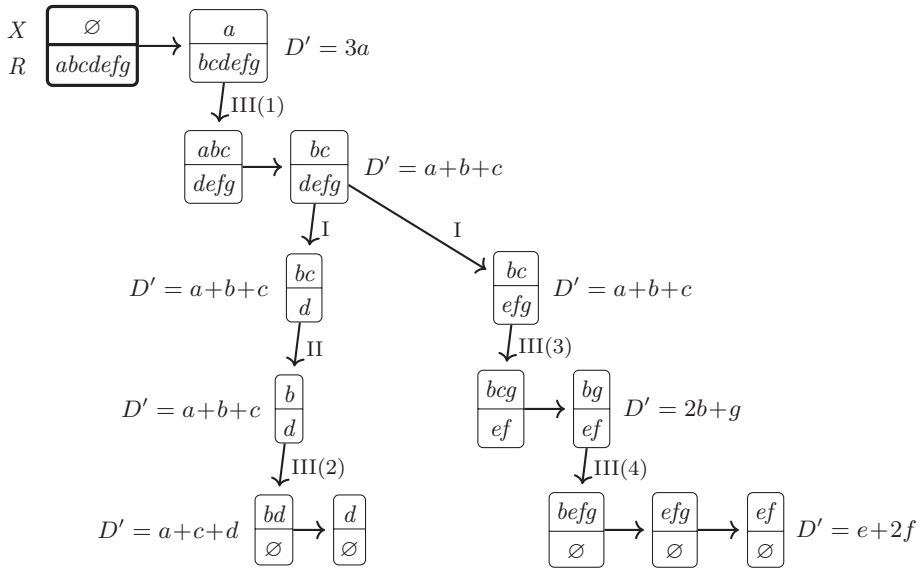


Figure 3.2: The monotone search strategy obtained from G with divisor $D = 3a$. Each node shows the corresponding pair (X, R) with the root being $(\varnothing, \{a, b, c, d, e, f, g\})$. The labels I–III refer to the steps in the construction.

If we follow the construction of §3.4, we will end up with the monotone search strategy found in Figure 3.2. Recall that every node of the search tree is a pair (X, R) of subsets of V such that R is a union of X-flaps in G. For every node, the sets X and R are indicated in the figure. The three ways of growing the tree (steps I, II, III of the construction) at a node with $R \neq \varnothing$ are indicated by downward arrows. Steps of type I are the only steps that involve branching of the tree. Steps of type III are the most involved: a path of nodes is added to the tree (depicted horizontally) and the divisor D' changes. For reference, the four steps of type III are labelled (1)–(4). Below, we will elaborate on the construction.

Root. The initial partial MSS consists of a root $(X, R) = (\varnothing, V)$ connected to a leaf

node $(\mathrm{supp}(D), V \setminus \mathrm{supp}(D)) = (\{a\}, \{b, c, d, e, f, g\})$. The divisor associated to the leaf node is $D' = 3a$.

Step III(1). For the leaf node $(X, R) = (\{a\}, \{b, c, d, e, f, g\})$, the set R is a single X-flap and $N(R) = X$, so we apply step III. Dhar's burning algorithm applied to divisor $D' = 3a$ (and an arbitrary vertex in R) gives the set $U = \{a\}$ to fire on. We have $U \cap X = \{s_1\}$ with $s_1 = a$. We obtain

$$X_1 = X \cup (N(s_1) \cap R) = \{a, b, c\}, \quad R_1 = R \setminus X_1 = \{d, e, f, g\}, \quad X_1' = \{b, c\}.$$

The path $(X_1, R_1) \to (X_1', R_1)$ is attached to leaf node (X, R). The divisor $D' - L_G \mathbb{1}_U = a + b + c$ is associated to the new leaf node (X_1', R_1).

Step I. For the leaf node $(X, R) = (\{b, c\}, \{defg\})$, the set R is the union of two X-flaps: $R = \{d\} \cup \{e, f, g\}$. Hence, we apply step I to obtain two new leaf nodes $(\{b, c\}, \{d\})$ and $(\{b, c\}, \{e, f, g\}$.

Step II. In the left-hand leaf node $(X, R) = (\{b, c\}, \{d\})$ we have $N(R) = \{b\} \subset X$, so we apply step II and add a new leaf node $(\{b\}, \{d\})$.

Step III(2–4). The remaining steps in the construction of the MSS are of type III. We summarize the details below.

(2) Divisor D' is equal to $a + b + c$. Applying Dhar's algorithm to D' and the vertex d, we obtain the set $U = \{a, b, c, e, f, g\}$. Firing on U, we obtain the new divisor $a + c + d$.

(3) Divisor D' is equal to $a + b + c$. Applying Dhar's algorithm to D' and any of the vertices in $R = \{e, f, g\}$ we obtain the set $U = \{a, c\}$. Firing on U, we obtain the new divisor $2b + g$.

(4) Divisor D' is equal to $2b + g$. Applying Dhar's algorithm to D' and the vertex e (or equivalently f) we obtain the set $U = \{a, b, c, d, g\}$. Firing on U, we obtain the new divisor $e + 2f$.

From the search tree, we obtain a tree-decomposition of width $\deg(D) = 3$ by labelling each node (X, R) by the set X and ignoring arc directions. Removing nodes with label $\varnothing$ and contracting edges of the tree between nodes with equal labels, we obtain the tree decomposition depicted in Figure 3.3.

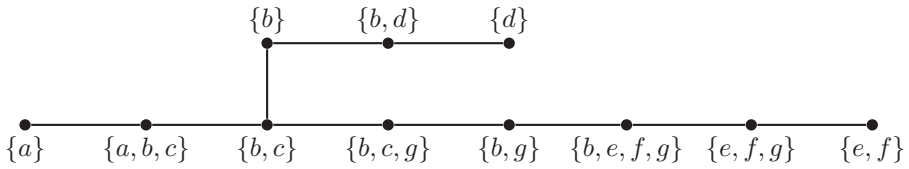


Figure 3.3: Tree decomposition of G derived from the MSS.

3.7 Closing remarks

We conclude this chapter with a few closing remarks.

First, as we mentioned in Chapter 2, there are several different (inequivalent) notions of gonality for graphs. In Part I of this dissertation, we focus on divisorial gonality. However, in the paper [BDGS22] upon which the present chapter is based, we also considered the so-called *stable gonality*, defined as the minimum degree of a finite harmonic morphism from a refinement of G to a tree (for details, see [BDGS22]). There we proved the following analogue of Theorem 3.1:

Theorem 3.11 ([BDGS22, Thm. 3])**.** *There is an $O(k^2 \cdot |V(G')|)$ time algorithm that takes as input a graph G and a finite harmonic morphism $f : G' \to T$ from a refinement G' of G to a tree T, and returns as output a tree decomposition of G of width at most k.*

The proof of Theorem 3.11 is relatively easy, as the harmonic morphism $f : G' \to T$ immediately gives rise to a tree decomposition (for details, see [BDGS22]). This also furnishes the first direct proof of the inequality $\mathrm{tw}(G) \leq \mathrm{sgon}(G)$, without relying on divisor theory.

Second, we point out that the results from this chapter have an application in parametrized complexity. It is well-known that many NP-hard graph problems become tractable for graphs of bounded treewidth, provided that a sufficiently small tree decomposition is available (see for instance [CFK+15, Thm. 7.9 and 7.10]). An immediate corollary of Theorem 3.1 is that the same NP-hard problems also become tractable when a positive rank divisor of sufficiently small degree is available. However, this could also be deduced without using Theorem 3.1, because one could simply ignore the divisor and use one of the existing algorithms for computing a tree decomposition, such as Bodlaender's celebrated $O(2^{O(k^3)} \cdot |V(G)|)$ time algorithm [Bod96]. In this setting, the benefit of Theorem 3.1 is that the dependence on k is linear instead of exponential (at the cost of higher dependence on $|V(G)|$ and $|E(G)|$), making the algorithm more practical for larger values of k.

Finally, we point out another intriguing connection with parametrized complexity. As mentioned in the previous paragraph, problems that are tractable when parametrized by treewidth are also tractable when parametrized by gonality. One important difference between treewidth and gonality is that treewidth is blind to parallel edges, whereas gonality is not. Motivated by this, Bodlaender, Cornelissen and van der Wegen were able to show that several well-studied classes of multigraph problems are hard when parametrized by treewidth, but become tractable when parametrized by stable gonality [BCW22b]. This suggests that gonality could play an important role in parametrized complexity of multigraph problems. An interesting open question is to also find problems which are hard when parametrized by treewidth but become tractable when parametrized by *divisorial* gonality [BCW22a, Open problem 5].

Chapter 4

Discrete and metric divisorial gonality can be different

In 2008, Matt Baker conjectured that the divisorial gonality of every graph G is equal to the divisorial gonality of every regular subdivision of G, and to the gonality of the associated metric graph $\Gamma(G, \mathbb{1})$ with unit lengths. In this chapter, we show that these two conjectures are equivalent, and we give a counterexample to both.

This chapter is based on the paper [DSW22], and is joint work with Harry Smit and Marieke van der Wegen.

4.1 Introduction

In [Bak08], Matt Baker provided a way to translate between curves and graphs, and used this to show that certain results can be carried over from one world to another. However, this translation is not perfect, as some information is lost in the process, so not all results from algebraic geometry could immediately be translated to analogous statements for graphs. As a consequence, the paper [Bak08] not only contains many new results, but also a number of conjectures. Since then, these conjectures have been among the main driving forces for further research into divisors and gonality on graphs.

All but two of the conjectures from [Bak08] have since been solved; see [HKN13, Luo11, CDPR12, DV21a]. The first and most important remaining open problem is the Brill–Noether conjecture for finite graphs, based on an analogous result for curves.

Conjecture 4.1 (Brill–Noether conjecture for graphs, [Bak08, Conj. 3.9])**.** *Define the* cyclomatic number *of a connected loopless multigraph G as $g := |E(G)| - |V(G)| + 1$. Furthermore, for integers $g, r, d \geq 0$, define the* Brill–Noether number *as*

$$\rho(g, r, d) := g - (r + 1)(g - d + r).$$

Then:

(a) If $\rho(g, r, d) \geq 0$, then every connected loopless multigraph G of cyclomatic number g has a divisor D with $\operatorname{rank}(D) = r$ *and* $\deg(D) \leq d$*;*

(b) If $\rho(g,r,d) < 0$, then there exists a connected loopless multigraph G of cyclomatic number g for which there is no divisor D with $\operatorname{rank}(D) = r$ *and* $\deg(D) \leq d$.

Parts (a) and (b) of Conjecture 4.1 are sometimes referred to as the 'existence' and 'non-existence' parts of the Brill–Noether conjecture, respectively (see e.g. [CDPR12, AR18, Man22]). The non-existence part (Conjecture 4.1(b)) was settled in 2012 by Cools, Draisma, Payne and Robeva [CDPR12], but the existence part (Conjecture 4.1(a)) remains wide open. We focus on the $r = 1$ case of this conjecture, which can be reformulated as follows.

Conjecture 4.2 (Gonality conjecture, [Bak08, Conjecture 3.10(1)])**.** *Let G be a connected loopless multigraph, and let $g := |E(G)| - |V(G)| + 1$ denote its cyclomatic number. Then* $\operatorname{dgon}(G) \leq \lfloor \frac{g+3}{2} \rfloor$.

The corresponding result for metric graphs was proved by Baker [Bak08, Thm. 3.12] using algebraic geometry. A purely combinatorial proof of this result was recently found by Draisma and Vargas [DV21a], with many promising avenues still to be explored [DV21b]. However, for discrete graphs, Conjecture 4.2 is still wide open.[1] Partial results were obtained by Atanasov and Ranganathan [AR18], who proved Conjecture 4.2 for all graphs of genus at most 5, and by Aidun and Morrison [AM20], who proved the conjecture for Cartesian product graphs.

The most straightforward approach to Conjecture 4.2 would be to show that the divisorial gonality of a graph is equal to the divisorial gonality of the associated metric graph with unit lengths (see §4.2). This is the second remaining conjecture of Baker's paper [Bak08, Conj. 3.14]. Given a multigraph G and an integer $k \geq 1$, let $\sigma_k(G)$ denote the multigraph obtained from G by subdividing every edge into k parts. The conjecture can then be stated as follows.

Conjecture 4.3 ([Bak08, Conjecture 3.14])**.** *Let G be a connected loopless multigraph, let $\Gamma(G)$ be the corresponding metric graph with unit edge lengths, and let $r \geq 1$. Then:*

(a) $\operatorname{dgon}_r(G) = \operatorname{dgon}_r(\sigma_k(G))$ *for all $k \geq 1$;*

(b) $\operatorname{dgon}_r(\Gamma(G)) = \operatorname{dgon}_r(G)$.

The first main result of this chapter is that Conjecture 4.3(a) and Conjecture 4.3(b) are equivalent for every graph G.

Theorem 4.4. *For every connected loopless multigraph G and every integer $r \geq 1$, one has*

$$\operatorname{dgon}_r(\Gamma(G)) = \min_{k \in \mathbb{N}_1} \operatorname{dgon}_r(\sigma_k(G)).$$

A partial result in this direction was already implicit in the work of Gathmann and Kerber [GK08, Prop. 3.1] (see Theorem 4.12(a) below), but to our knowledge Theorem 4.4 is new. Moreover, we use a different proof technique, which can be used

[1] A proof of Conjecture 4.1(a), and hence in particular Conjecture 4.2, was given by Caporaso in [Cap12, Thm. 6.3], but a gap in this proof was later pointed out by Sam Payne and reported by Baker and Jensen in [BJ16, Rmk. 4.8 and footnote 5]. To our knowledge, this has not been repaired.

to give an upper bound on the number of subdivisions needed to get equality (see Remark 4.16).

The proof runs roughly as follows. It is already known that every rank r divisor on $\sigma_k(G)$ also defines a rank r divisor on $\Gamma(G)$. For the converse, we show that every rank r divisor D on $\Gamma(G)$ can be "rounded" to a nearby divisor D' with $\operatorname{rank}(D') \geq r$ which is supported on the $\mathbb{Q}$-points of $\Gamma(G)$, and therefore on the points of some regular subdivision $\sigma_k(G)$. The details will be given in §4.3.

As pointed out by Baker in [Bak08], a positive answer to Conjecture 4.3 would also yield a positive answer to Conjecture 4.2. However, it turns out that the subdivision conjecture fails, and we give a counterexample to Conjecture 4.3(a) in the case $r = 1$ and $k = 2$. Evidently this is also a counterexample to Conjecture 4.3(b). The second main result of this chapter is the following.

Theorem 4.5. *For every integer $k \geq 1$, there exists a connected loopless multigraph G_k such that $\operatorname{dgon}(G_k) = 6k$ and $\operatorname{dgon}(\Gamma(G_k)) = \operatorname{dgon}(\sigma_2(G_k)) = 5k$. Furthermore, G_k can be chosen simple and bipartite.*

The proof is constructive and consists of two parts. In §4.4, we construct a family of graphs with $\operatorname{dgon}(G) = 6$ and $\operatorname{dgon}(\Gamma(G)) = \operatorname{dgon}(\sigma_2(G)) = 5$. The graphs G_k are then constructed in §4.5 by combining k of these graphs in a certain way.

Although the difference between $\operatorname{dgon}(G)$ and $\operatorname{dgon}(\Gamma(G))$ can be large, as in Theorem 4.5, the ratio between them is at most 2, as we show in Proposition 4.27. Hence, for the gap to get arbitrarily large, it is necessary that $\operatorname{dgon}(\Gamma(G))$ goes to infinity.

In §4.6, we list a few additional counterexamples (without proof), including a 3-regular graph. Although all counterexamples in this chapter violate Conjecture 4.3, they nevertheless satisfy the Brill–Noether bound. We do not know whether any of these examples can be extended to disprove Conjecture 4.2. Additional open problems are discussed in §4.6 as well.

4.2 Metric graphs and rank-determining sets

Before we dive into the proofs of Theorem 4.4 and Theorem 4.5, we briefly recall the basics of metric graphs and their divisor theory.

A *metric graph* is a metric space Γ that can be obtained in the following way. Let G be a finite multigraph and let $\ell\colon E(G) \to \mathbb{R}_{>0}$ be an assignment of lengths to the edges of G. To construct Γ, take an interval $[0, \ell(e)]$ for every edge $e \in E(G)$, and glue these together at the endpoints as prescribed by G. To turn it into a metric space, equip Γ with the shortest path metric in the obvious way. The metric graph Γ defined in this way will be denoted $\Gamma(G, \ell)$. If $\ell = \mathbb{1}$ is the unit length function, we write $\Gamma(G) := \Gamma(G, \mathbb{1})$.

If the metric graph Γ is constructed from the pair (G, ℓ) as above, then we say that (G, ℓ) is a *model* of Γ. We say that a model (G, ℓ) is *loopless* (resp. *simple*) if G is loopless (resp. simple). The *valency* $\operatorname{val}(v)$ of $v \in \Gamma$ is the number of edges incident with v in any loopless model (G, ℓ) with $v \in V(G)$.

A *divisor* on a metric graph Γ is an element of the free abelian group on Γ. In other words, a divisor is a formal sum $\sum_{v\in\Gamma} a_v v$ where $a_v \in \mathbb{Z}$ for all v, and $a_v = 0$ for all but finitely many v. The notations $\operatorname{supp}(D)$, $\deg(D)$, $D \geq D'$, $\operatorname{Div}(\Gamma)$, $\operatorname{Div}_+(\Gamma)$ and $\operatorname{Div}^d_+(\Gamma)$ are defined analogously to the discrete case.

The definition of equivalence is a bit different from the discrete case (see §2.2). A *rational function* on Γ is a continuous piecewise linear function $f\colon \Gamma \to \mathbb{R}$ with integral slopes. For each point $v \in \Gamma$, let a_v be the sum of the outgoing slopes of f in all edges incident with v in some appropriate model of Γ. The corresponding divisor $\sum_{v\in\Gamma} a_v v$ is called a *principal divisor*. Two divisors D and D' are *equivalent* if $D - D'$ is a principal divisor.

The *rank* of a divisor $D \in \operatorname{Div}(\Gamma)$ is defined as in the discrete case; that is:

$$\operatorname{rank}(D) := \max\{k \in \mathbb{Z} \mid D-E \text{ is equivalent to an effective divisor for all } E \in \operatorname{Div}^k_+(\Gamma)\}.$$

The *r-th (divisorial) gonality* $\operatorname{dgon}_r(\Gamma)$ of Γ is the minimum degree of a rank r divisor on Γ. For $r = 1$, this is simply called the *(divisorial) gonality* of Γ: $\operatorname{dgon}(\Gamma) := \operatorname{dgon}_1(\Gamma)$.

If G is a finite graph and if $\Gamma := \Gamma(G)$ is the corresponding metric graph with unit lengths, then two divisors $D, D' \in \operatorname{Div}(G)$ are equivalent on G if and only if they are equivalent on Γ; see [Bak08, Rmk. 1.3]. Furthermore, in this case one has $\operatorname{rank}_G(D) = \operatorname{rank}_\Gamma(D)$ for every divisor $D \in \operatorname{Div}(G)$; see [HKN13, Thm. 1.3].

Let Γ be a metric graph, and let $S \subseteq \Gamma$ be a subset. Following [Luo11], we define the *S-restricted rank* of a divisor $D \in \operatorname{Div}(\Gamma)$ as

$$\operatorname{rank}_S(D) := \max\{k \mid D-E \text{ is equivalent to an effective divisor for all } E \in \operatorname{Div}^k_+(S)\},$$

where $\operatorname{Div}^k_+(S)$ is the set of degree k effective divisors whose support is contained in S. The set S is *rank-determining* if $\operatorname{rank}_S(D) = \operatorname{rank}(D)$ for all $D \in \operatorname{Div}(\Gamma)$. The following theorems are due to Luo.

Theorem 4.6 ([Luo11, Thm. 1.6]; see also [HKN13, Thm. 1.7])**.** *Let Γ be a metric graph, and let (G, ℓ) be a loopless model of Γ. Then the set $V(G) \subseteq \Gamma$ is rank-determining.*

Theorem 4.7 ([Luo11, Thm. 1.10])**.** *Let Γ, Γ' be metric graphs, and let $\phi\colon \Gamma \to \Gamma'$ be a homeomorphism. Then $S \subseteq \Gamma$ is rank-determining if and only if $\phi[S] \subseteq \Gamma'$ is rank-determining.*

We also formulate a discrete analogue of Theorem 4.6 for the case $r = 1$. If G is a graph, then we say that a divisor $D \in \operatorname{Div}(G)$ *reaches* the vertex $v \in V(G)$ if there is an effective divisor D' equivalent to D with $D'(v) > 0$. Furthermore, we say that a set $S \subseteq V(G)$ is a *strong separator* if for every connected component C of $V(G) \setminus S$ we have that C is a tree and for every $s \in S$ there is at most one edge (in G) between C and s.

Theorem 4.8 ([DG20, Lem. 2.6])**.** *Let G be a graph, and let $S \subseteq V(G)$ be a strong separator. If $D \in \operatorname{Div}(G)$ reaches every $s \in S$, then* $\operatorname{rank}(D) \geq 1$.

The following corollary is immediate from either Theorem 4.6 or Theorem 4.8.

Corollary 4.9. *Let G be a loopless multigraph, and let H be a subdivision of G. If $D \in \operatorname{Div}(H)$ reaches all vertices of $V(G)$, then* $\operatorname{rank}(D) \geq 1$.

4.3 Equivalence of the subdivision conjecture and the metrization conjecture

In this section, we prove Theorem 4.4 using a modification of the proof of [DG20, Thm. 5.1]. The main idea is the following: given a rank r divisor D on the metric graph $\Gamma(G)$, we will change the lengths of the edges between points in $V(G) \cup \operatorname{supp}(D)$ in such a way that $\operatorname{supp}(D)$ is moved to the $\mathbb{Q}$-points of the graph, all the while leaving the rank of D and the distances between the vertices of G unchanged. We will now make this precise.

Definition 4.10. Given a metric graph Γ and a model (G, ℓ) of Γ, a *G-rescaling of* Γ is a metric graph $\Gamma' := \Gamma(G, \ell')$, where $\ell' \in \mathbb{R}_{>0}^{E(G)}$ is another length vector. If $D \in \operatorname{Div}(\Gamma)$ with $\operatorname{supp}(D) \subseteq V(G)$, then D defines a divisor $D' \in \operatorname{Div}(\Gamma')$ in the obvious way, which we call *the G-rescaling of D*.

We point out that Γ and its G-rescaling Γ' can be isometric even if $\ell \neq \ell'$. This is because vertices of degree 2 can be moved around, as illustrated in Figure 4.1. In that case the vertex set $V(G)$ is embedded into $\Gamma \cong \Gamma'$ in two different ways, and the divisor D and its G-rescaling D' could be different divisors on the same metric graph. This will be the main tool in our proof of Theorem 4.4.

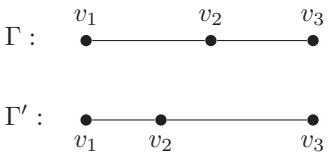


Figure 4.1: A metric graph $\Gamma = \Gamma(G, \ell)$ and a rescaling $\Gamma' = \Gamma(G, \ell')$ with $\ell' \neq \ell$ such that Γ and Γ' are isometric.

To rescale from real to rational edge lengths we use the following lemma.

Lemma 4.11. *Let $A \in \mathbb{Q}^{m \times n}$ and $b \in \mathbb{Q}^m$. If the linear system $Ax = b$ has a solution $x \in \mathbb{R}_{>0}^n$, then it also has a solution $x' \in \mathbb{Q}_{>0}^n$.*

Proof. Since the system has a solution $x \in \mathbb{R}_{>0}$, the solution space $\{z \mid Az = b\}$ is a non-empty affine $\mathbb{Q}$-subspace of $\mathbb{R}^n$. Choose an affine rational basis $y_0, \ldots, y_d \in \mathbb{Q}^n$ for the solution space and write $x = \alpha^{(0)} y_0 + \cdots + \alpha^{(d)} y_d$ with $\alpha^{(0)} + \cdots + \alpha^{(d)} = 1$. For every $i \in \{1, \ldots, d\}$, choose a rational sequence $\{\alpha_k^{(i)}\}_{k=1}^{\infty}$ such that $\lim_{k\to\infty} \alpha_k^{(i)} = \alpha^{(i)}$, and define $\alpha_k^{(0)} := 1 - \alpha_k^{(1)} - \cdots - \alpha_k^{(d)}$. Then $\lim_{k\to\infty} \alpha_k^{(0)} y_0 + \cdots + \alpha_k^{(d)} y_d = x$. Since $\mathbb{R}_{>0}^n$ is an open neighbourhood of x, there is a $K_0 \in \mathbb{N}$ such that $\alpha_k^{(0)} y_0 + \cdots + \alpha_k^{(d)} y_d \in \mathbb{R}_{>0}^n$ for all $k \geq K_0$. This gives a sequence of solutions in $\mathbb{Q}_{>0}^n$ converging to x. Any one of these suffices. ∎

We now come to the main result of this section. This is an extension of [DG20, Thm. 5.1], though the result of Theorem 4.12(a) was already implicit in the proof of [GK08, Prop. 3.1].

Theorem 4.12. *Let Γ be a metric graph, and let $D \in \mathrm{Div}_+(\Gamma)$ be an effective divisor.*

(a) There exists a loopless model (G, ℓ) with $\mathrm{supp}(D) \subseteq V(G)$ and a rational length vector $\ell' \in \mathbb{Q}_{>0}^{E(G)}$ such that the G-rescaling D' of D in $\Gamma' := \Gamma(G, \ell')$ satisfies $\mathrm{rank}_{\Gamma'}(D') \geq \mathrm{rank}_\Gamma(D)$.

(b) If Γ is a metric $\mathbb{Q}$-graph, then the length vector ℓ' in (a) can be chosen in such a way that Γ' is isometric to Γ.

Proof. (a) Write $r := \mathrm{rank}_\Gamma(D)$, and let $S \subseteq \Gamma$ be a finite rank-determining set. For every $E \in \mathrm{Div}_+^r(S)$, choose a divisor $D_E \in \mathrm{Div}(\Gamma)$ and a rational function $f_E \colon \Gamma \to \mathbb{R}$ such that $D_E \geq E$ and $D_E = D + \mathrm{div}(f_E)$. Furthermore, choose a loopless model (G, ℓ) of Γ such that

$$S \cup \mathrm{supp}(D) \cup \bigcup_{E \in \mathrm{Div}_+^r(S)} \mathrm{supp}(D_E) \subseteq V(G).$$

Since $D, D_E \geq 0$ and $D_E - D = \mathrm{div}(f_E)$, we have $\mathrm{supp}(\mathrm{div}(f_E)) \subseteq \mathrm{supp}(D) \cup \mathrm{supp}(D_E) \subseteq V(G)$, so $V(G)$ contains all points of non-linearity of f_E, for every $E \in \mathrm{Div}_+^r(S)$.

Choose an orientation of the edges of G. For every cycle C in G, choose a circular orientation of the edges of C, and define $\chi_C \colon E(G) \to \{-1, 0, 1\}$ by setting

$$\chi_C(e) := \begin{cases} 1, & \text{if } e \in E(C) \text{ and the orientations of } G \text{ and } C \text{ agree on } e; \\ -1, & \text{if } e \in E(C) \text{ and the orientations of } G \text{ and } C \text{ disagree on } e; \\ 0, & \text{if } e \notin E(C). \end{cases}$$

For $E \in \mathrm{Div}_+^r(S)$ and $e \in E(G)$, let $\phi(f_E, e) \in \mathbb{Z}$ denote the slope of f_E on e, in the forward direction of e. Note that a G-rescaling $\Gamma(G, \ell')$ of Γ admits rational functions f'_E whose slope on e equals $\phi(f_E, e)$, for all $E \in \mathrm{Div}_+^r(S)$ and all $e \in E(G)$, if and only if $y = \ell'$ is a solution to following system of equations:

$$\sum_{e \in E(G)} \phi(f_E, e) \chi_C(e)\, y(e) = 0, \text{ for every cycle } C \text{ and every } E \in \mathrm{Div}_+^r(S). \tag{4.13}$$

Since the coefficients (that is, $\phi(f_E, e)\chi_C(e)$) and constants (that is, 0) of this linear system are integral, and since $y = \ell \in \mathbb{R}_{>0}^{E(G)}$ is a solution, it follows from Lemma 4.11 that there exists a solution $\ell' \in \mathbb{Q}_{>0}^{E(G)}$.

Consider the G-rescaling $\Gamma' := \Gamma(G, \ell')$. Let D' be the corresponding G-rescaling of D, and let D'_E be the G-rescaling of D_E for all $E \in \mathrm{Div}_+^r(S)$. By the above, we may choose rational functions f'_E on Γ' such that the slope of f'_E on e equals $\phi(f_E, e)$, for all $E \in \mathrm{Div}_+^r(S)$ and all $e \in E(G)$. Then clearly $D'_E = D' + \mathrm{div}(f'_E)$, so the D'_E are equivalent to D'. Since $D'_E \geq E$ for every $E \in \mathrm{Div}_+^r(S)$, it follows that $\mathrm{rank}_S(D') \geq r$. By Theorem 4.7, S is rank-determining in Γ', so $\mathrm{rank}_{\Gamma'}(D') = \mathrm{rank}_S(D') \geq r$.

(b) Choose a rational model $(\tilde{G}, \tilde{\ell})$ of Γ. We repeat the argument of (a) with the following modifications. First, we add the requirement that $V(\tilde{G}) \subseteq V(G)$. Then every edge $\tilde{e}$ in $\tilde{G}$ corresponds to a path in G, which we denote $P_{\tilde{e}}$. Second, we extend the linear system from (4.13) by adding the following constraints:

$$\sum_{e \in E(P_{\tilde{e}})} y(e) = \tilde{\ell}(\tilde{e}), \quad \text{for all } \tilde{e} \in E(\tilde{G}). \tag{4.14}$$

Again, the coefficients and constants of the linear system are rational, and $y = \ell \in \mathbb{R}_{>0}^{E(G)}$ is a solution, so it follows from Lemma 4.11 that there is a solution $\ell' \in \mathbb{Q}_{>0}^{E(G)}$. The rest of the proof of (a) carries through unchanged, and the extra constraints from (4.14) ensure that Γ' is isometric to Γ. ■

Proof of Theorem 4.4. A rank r divisor on $\sigma_k(G)$ also defines a rank r divisor on $\Gamma(\sigma_k(G), \mathbb{1}/k) = \Gamma(G, \mathbb{1})$. Therefore $\operatorname{dgon}_r(\Gamma(G)) \leq \min_{k \in \mathbb{N}_1} \operatorname{dgon}_r(\sigma_k(G))$.

Conversely, let $D \in \operatorname{Div}_+(\Gamma(G))$ be an effective divisor of rank r. By Theorem 4.12(b), there exists a divisor $D' \in \operatorname{Div}_+(\Gamma(G))$ with $\deg(D') = \deg(D)$ and $\operatorname{rank}(D') \geq \operatorname{rank}(D)$ which is supported on the $\mathbb{Q}$-points of $\Gamma(G)$. Then D' is supported on the vertices of the model $(\sigma_k(G), \mathbb{1}/k)$ of $\Gamma(G)$ for some $k \in \mathbb{N}_1$, so we have $\operatorname{dgon}_r(\sigma_k(G)) \leq \operatorname{dgon}_r(\Gamma(G))$. ■

Remark 4.15. Analogously to the proof of the main result of [BWZ21], the linear system from the proof of Theorem 4.12(b) forms a certificate that $\operatorname{dgon}_r(\Gamma) \leq d$. If r and d are fixed, then this certificate has size polynomial in the size of Γ, so it follows that Metric Divisorial r-Gonality for $\mathbb{Q}$-graphs belongs to the complexity class NP. (For details, refer to the proof in [BWZ21].) This problem is also known to be NP-hard: for $r = 1$, this can be deduced from the proof of [GSW20, Thm. 3.5] (see also [EEH+22, Thm. 1.3], where a different proof is given), and for arbitrary $r \geq 1$ this was proved in [MT22].

Remark 4.16. The proof of Theorem 4.12(b) can also be used to find an upper bound on the size of the subdivision needed to get equality in Theorem 4.4. One such upper bound can be obtained by following the proof of [BWZ21, Cor. 6.2]. We sketch a way to improve this bound. Let $\tilde{G}$ be a graph with n vertices and m edges, let $\Gamma := \Gamma(\tilde{G})$ be the corresponding unit metric graph, and let $D \in \operatorname{Div}(\Gamma)$ be a divisor of degree d and rank r. We repeat the proof of Theorem 4.12(b) with respect to the rational model $(\tilde{G}, \mathbb{1})$ and the rank-determining set $S := V(\tilde{G})$ (use Theorem 4.6). Without loss of generality, we may assume that D is equal to one of the D_E. Then the number of variables of the linear system is $|E(G)| \leq m + dn^r$.

Note that we can also allow a solution $\ell' \geq 0$ instead of $\ell' > 0$. This has the effect of contracting some of the edges of the model G from the proof of Theorem 4.12, but the equations from (4.14) ensure that the resulting graph Γ' is still isometric to Γ. Hence (4.13) and (4.14) determine a linear program $Ax = b$, $x \geq 0$, and the entries of A are integers which can be shown to be bounded in absolute value by d. The set of feasible solutions is non-empty and bounded by (4.14), so there is a basic feasible solution x (see e.g. [MG07, Thm. 4.2.3]). Hence there is a subset $B \subseteq \{1, \ldots, |E(G)|\}$ such that $x_B = A_B^{-1} b$ and $x_{B^c} = 0$. Therefore the lowest common denominator of the entries

of x is at most $|\det(A_B)| \leq \sum_{\sigma \in \mathrm{Sym}(B)} \prod_{i \in B} |A_{i\sigma(i)}| \leq |B|! \cdot d^{|B|}$. In conclusion, if the unit metric graph $\Gamma = \Gamma(\tilde{G}, \mathbb{1})$ has a divisor of rank r and degree d, then so does $\sigma_k(\tilde{G})$ for some $k \leq (m + dn^r)! \cdot d^{m+dn^r}$.

4.4 A graph G such that $\mathrm{dgon}(\sigma_2(G)) < \mathrm{dgon}(G)$

In this section we construct a class of graphs, which we call "tricycle graphs". We show that the divisorial gonality of any tricycle graph G is strictly greater than the divisorial gonality of its 2-subdivision $\sigma_2(G)$, and thus of its associated metric graph $\Gamma(G)$.

Definition 4.17. A *tricycle graph* is a multigraph G that can be obtained in the following way:

- Start with three disjoint cycles C_1, C_2, C_3, each on at least 2 vertices (a cycle on 2 vertices consists of two vertices connected by two parallel edges).
- Choose two distinct vertices on each of these cycles, say $v_i^-, v_i^+ \in V(C_i)$.
- Add the edges $v_1^+v_2^-$, $v_2^+v_3^-$ and $v_3^+v_1^-$ to the graph.
- Add another vertex v_0 and six new edges to G, connecting v_0 to the six vertices $v_1^-, v_1^+, v_2^-, v_2^+, v_3^-, v_3^+$.
- Subdivide the six edges incident with v_0.

We call the vertices $v_i^-, v_i^+ \in V(C_i)$ the *transition vertices*, the edges $v_1^+v_2^-$, $v_2^+v_3^-$ and $v_3^+v_1^-$ the *transition edges*, and v_0 the *central vertex*. The *outer ring* is the union of the cycles C_1, C_2, C_3 and the transition edges.

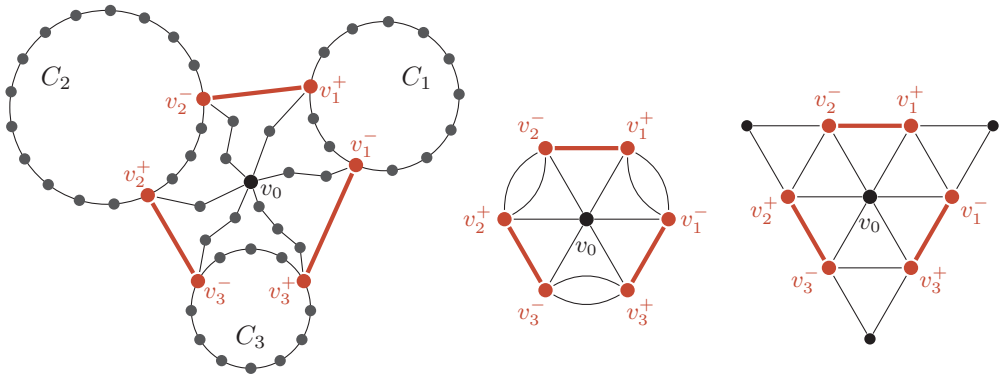
Figure 4.2 illustrates an example of a tricycle graph, along with the minimal tricycle T_{m} and the minimal simple tricycle T_{ms}. Note that a multigraph (resp. simple graph) G is a tricycle if and only if G can be obtained by taking a subdivision of the minimal tricycle T_{m} (resp. the minimal simple tricycle T_{ms}) in such a way that the transition edges are not subdivided.

In what follows, we will show that every tricycle graph G satisfies $\mathrm{dgon}(G) = 6$ and $\mathrm{dgon}(\Gamma(G)) = \mathrm{dgon}(\sigma_2(G)) = 5$. First of all, we exhibit a positive rank divisor of degree 5 on $\sigma_2(G)$.

Proposition 4.18. *Let G be a tricycle graph. Then* $\mathrm{dgon}(\sigma_2(G)) \leq 5$.

Proof. Let $D_0 \in \mathrm{Div}(\sigma_2(G))$ be the effective divisor with two chips on v_0 and one chip on the midpoint of each of the transition edges $v_1^+v_2^-$, $v_2^+v_3^-$, $v_3^+v_1^-$. Then $\deg(D_0) = 5$. In light of Corollary 4.9, in order to show that D_0 has positive rank, it suffices to prove that D_0 reaches v_0 and the transition vertices $v_1^-, v_1^+, v_2^-, v_2^+, v_3^-, v_3^+$.

Clearly D_0 reaches v_0, for we have $D_0(v_0) > 0$. Now fix $i \in \{1, 2, 3\}$. To reach the transition vertices v_i^- and v_i^+, let $S_i \subseteq V(\sigma_2(G))$ be the connected component of $\sigma_2(G) \setminus \mathrm{supp}(D_0)$ that contains the cycle C_i. Then the subset S_i^c can be fired, and doing so yields an effective divisor D_i with $D_i(v_i^-) = D_i(v_i^+) = 1$. This shows that D_0



(a) a generic tricycle graph; (b) the minimal tricycle T_{m}; (c) the minimal simple tricycle T_{ms}.

Figure 4.2: A generic tricycle, the minimal tricycle, and the minimal simple tricycle, with the transition vertices and the transition edges highlighted for emphasis.

reaches v_i^- and v_i^+, for all $i \in \{1, 2, 3\}$. It follows that D_0 is a positive rank divisor on $\sigma_2(G)$, hence $\operatorname{dgon}(\sigma_2(G)) \leq 5$. ■

Evidently, the divisor from Proposition 4.18 is not supported on vertices of G. The remainder of this section is dedicated to showing that G has no positive rank divisors of degree 5. Along the way, we also prove that $\operatorname{dgon}(\Gamma(G)) \geq 5$.

In Lemma 4.21 below, we show that every positive rank v_0-reduced divisor of degree at most 5 on a subdivision of the minimal tricycle T_{m} must be of a very specific form. This will subsequently be used to show that $\operatorname{dgon}(\Gamma(G)) = \operatorname{dgon}(\sigma_2(G)) = 5$ (see Corollary 4.23) and $\operatorname{dgon}(G) = 6$ (see Theorem 4.24) for every tricycle graph G.

For convenience, we use the following notation.

Definition 4.19. Let G be a graph and let H be a subdivision of G. For $e = uw \in E(G)$, let $P^e_{[u,w]} \subseteq H$ denote the path $uv_1v_2 \cdots v_k w$ in H corresponding to the subdivided edge e. Furthermore, let $P^e_{(u,w)} := P^e_{[u,w]} \setminus \{u, w\}$, $P^e_{[u,w)} := P^e_{[u,w]} \setminus \{w\}$ and $P^e_{(u,w]} := P^e_{[u,w]} \setminus \{u\}$ denote the corresponding open and half-open subpaths. If e is the only edge between u and w, then we omit the superscript and simply write $P_{[u,w]}$, $P_{(u,w)}$, $P_{[u,w)}$ and $P_{(u,w]}$.

The following simple lemma is essential to our proof, and will be used repeatedly.

Lemma 4.20. *Let G be a graph and let $v_0 \in V(G)$. Let $e_1, \ldots, e_k \in V(G)$ be the edges incident with v_0, and let $v_i \in V(G) \setminus \{v_0\}$ be the other endpoint of e_i for every i. Moreover, let H be a subdivision of G, let $D \in \operatorname{Div}(H)$ be a positive rank v_0-reduced divisor on H, and let $w \in V(H)$ be a vertex with $D(w) = 0$.*

Then an execution of Dhar's burning algorithm on the triple (H, D, w) has the following properties:

(a) v_0 is not burned;

(b) If $I \subseteq \{i \in [k] : v_i$ is burned$\}$, then $\bigcup_{i\in I} P^{e_i}_{[v_0,v_i)}$ contains at least $|I|$ chips.

Proof. (a) Since D has positive rank and $D(w) = 0$, the divisor D cannot be w-reduced, so Dhar's algorithm returns a non-empty subset $U \subseteq V(G)$ that can be fired. Since D is v_0-reduced, we must have $v_0 \in U$, which means that v_0 is not burned.

(b) Partition I as $I = I_0 \cup I_1$, where $i \in I_0$ if all vertices of the path $P^{e_i}_{(v_0,v_i]}$ are burned, and $i \in I_1$ otherwise. Since v_0 is not burned, it has at most $D(v_0)$ burning neighbours, so $|I_0| \leq D(v_0)$. Moreover, if $i \in I_1$, then v_i is burned, but not all vertices of the path $P^{e_i}_{(v_0,v_i]}$ are burned, so there must be at least one chip on $P^{e_i}_{(v_0,v_i)}$. The conclusion follows. ■

We will apply Lemma 4.20 to an arbitrary subdivision of the minimal tricycle T_{m}. For this we use the following terminology. Using notation from Definition 4.19, if H is a subdivision of T_{m}, then the three transition edges $v_1^+v_2^-$, $v_2^+v_3^-$, $v_3^+v_1^-$ of T_{m} correspond to the paths $P_{[v_1^+,v_2^-]}$, $P_{[v_2^+,v_3^-]}$, $P_{[v_3^+,v_1^-]}$ in H, which we call the *transition paths*. The *transition vertices* of H are the images in H of the original six transition vertices $v_1^-, v_1^+, v_2^-, v_2^+, v_3^-, v_3^+$ of T_{m}, or in other words, the endpoints of the transition paths in H. (This is consistent with our definition of the transition vertices of a tricycle graph, which can also be seen as a subdivision of T_{m}.)

Lemma 4.21. *Let H be a subdivision of the minimal tricycle T_{m}. If $D \in \mathrm{Div}(H)$ is a positive rank v_0-reduced divisor with $\deg(D) \leq 5$, then D must have two chips on v_0 and exactly one chip on each of the transition paths $P_{[v_1^+,v_2^-]}$, $P_{[v_2^+,v_3^-]}$, $P_{[v_3^+,v_1^-]}$.*

Proof. First, we prove that there must be at least one chip on every transition path. Suppose, for the sake of contradiction, that one of the transition paths, say $P_{[v_1^+,v_2^-]}$, has no chips at all. We start an execution of Dhar's burning algorithm on (H, D, v_1^+). Let $H_2^+ \subseteq H$ be the union of the cycle C_2 and the transition path $P_{[v_2^+,v_3^-]}$.

We claim that the number of chips on H_2^+ plus the number of burned transition vertices in H_2^+ is at least 3. To that end, note first of all that v_2^- is burned, since there is no chip on the transition path $P_{[v_1^+,v_2^-]}$. Now we distinguish three cases:

- If v_2^+ is not burned, then there must be at least two chips on C_2 to stop the fire spreading from v_2^- to v_2^+. In this case, H_2^+ contains at least one burned transition vertex (namely v_2^-) and at least two chips, for a total of at least 3.

- If v_2^+ is burned but v_3^- is not burned, then there must be at least one chip on the half-open transition path $P_{(v_2^+,v_3^-]}$. In this case, H_2^+ contains two burned transition vertices (v_2^- and v_2^+) and at least one chip, for a total of at least 3;

- If both v_2^+ and v_3^- are burned, then H_2^+ contains three burned transition vertices (v_2^-, v_2^+ and v_3^-).

Likewise, write $H_1^- := C_1 \cup P_{[v_1^-, v_3^+]}$. Analogously, the number of chips plus the number of burned transition vertices on H_1^- is at least 3. Since H_2^+ and H_1^- are disjoint, the total number of chips on the outer ring plus the total number of burned transition vertices is at least 6. But since the transition vertices are exactly the T_{m}-neighbours of v_0, and since the half-open paths $P_{[v_0, v_i^\pm)}$ are disjoint from the outer ring, it follows from Lemma 4.20(b) that $\deg(D) \geq 6$, which is a contradiction. We conclude that every transition path must have at least one chip.

Second, we prove that there must be two chips on v_0. Since the total number of chips is at most 5, there must be a cycle C_i on the outer ring with at most one chip. Choose $w \in V(C_i)$ with $D(w) = 0$ and start an execution of Dhar's burning algorithm on (H, D, w). Since there is at most one chip on C_i, the entire cycle C_i is burned. It follows from Lemma 4.20(b) that there are at least two chips on $P_{[v_0, v_i^-)} \cup P_{[v_0, v_i^+)}$. Therefore the number of chips on the outer ring is at most 3, so there must be another cycle C_j ($j \neq i$) on the outer ring with at most one chip. By an analogous argument, there are at least two chips on $P_{[v_0, v_j^+)} \cup P_{[v_0, v_j^-)}$. But since the outer ring has at least 3 chips (one on every transition path), there can be at most 2 chips on $P_{[v_0, v_i^-)} \cup P_{[v_0, v_i^+)} \cup P_{[v_0, v_j^-)} \cup P_{[v_0, v_j^+)}$. The only way to meet these requirements is if there are exactly two chips on v_0.

To conclude the proof, note that 2 chips on v_0 and at least 1 chip on every transition path add up to at least 5 chips in total. Since $\deg(D) \leq 5$, all chips have been accounted for. In particular, there cannot be more than one chip on each of the transition paths. ■

Lemma 4.21 shows that every positive rank v_0-reduced divisor D with $\deg(D) \leq 5$ must in fact satisfy $\deg(D) = 5$, so the following corollary is immediate.

Corollary 4.22. *Let H be a subdivision of the minimal tricycle T_{m}. Then one has* $\operatorname{dgon}(H) \geq 5$.

In particular, this enables us to compute the metric gonality of an arbitrary tricycle graph:

Corollary 4.23. *Let G be a tricycle graph. Then* $\operatorname{dgon}(\Gamma(G)) = \operatorname{dgon}(\sigma_2(G)) = 5$.

Proof. It follows from Proposition 4.18 that $\operatorname{dgon}(\Gamma(G)) \leq \operatorname{dgon}(\sigma_2(G)) \leq 5$. Furthermore, since every subdivision of G is a also subdivision of the minimal tricycle T_{m}, it follows from Corollary 4.22 and Theorem 4.4 that

$$\operatorname{dgon}(\Gamma(G)) = \min_{k \in \mathbb{N}_1} \operatorname{dgon}(\sigma_k(G)) \geq 5. \qquad \blacksquare$$

All that remains is to prove that every tricycle graph has divisorial gonality 6. To do so, we once again use the preceding lemmas.

Theorem 4.24. *Every tricycle graph G satisfies* $\operatorname{dgon}(G) = 6$.

Proof. Suppose, for the sake of contradiction, that $\operatorname{dgon}(G) \leq 5$. Then we may choose a positive rank v_0-reduced divisor $D \in \operatorname{Div}(G)$ with $\deg(D) \leq 5$. We interpret G as a subdivision of the minimal tricycle T_{m}. It follows from Lemma 4.21 that D has two chips on v_0 and exactly one chip on every transition path. Since G is a tricycle graph, the transition edges of T_{m} are not subdivided. Therefore a chip on a transition edge must lie on one of the transition vertices.

By the above, the divisor D has between 0 and 2 chips on each of the cycles C_1, C_2, C_3 on the outer ring, and all such chips must lie on the transition vertices. Since the total number of chips on the outer ring is odd, there must be a cycle C_i with exactly one chip. Assume without loss of generality that C_1 is such a cycle, and that $D(v_1^-) = 0$ and $D(v_1^+) = 1$.

We start an execution of Dhar's burning algorithm on (G, D, v_1^-). Since there is only one chip on C_1, the entire cycle C_1 is burned. In particular, the vertex v_1^+ is burned. The transition edge $v_1^+v_2^-$ has exactly one chip, which is on v_1^+, so the fire spreads via this edge to the vertex v_2^-, which is also burned. But now we see that at least three T_{m}-neighbours of v_0 are burned (namely, v_1^-, v_1^+ and v_2^-), so it follows from Lemma 4.20(b) that there must be at least 3 chips on $P_{[v_0,v_1^-)} \cup P_{[v_0,v_1^+)} \cup P_{[v_0,v_2^-)}$. This is a contradiction, and we conclude that $\operatorname{dgon}(G) \geq 6$.

To see that $\operatorname{dgon}(G) \leq 6$, note that the set $\{v_1^-, v_1^+, v_2^-, v_2^+, v_3^-, v_3^+\}$ of all transition vertices is a strong separator. Therefore the effective divisor with one chip on each of the transition vertices has positive rank, by Theorem 4.8, so $\operatorname{dgon}(G) \leq 6$. ■

This concludes the proof of validity of our counterexample. In summary, every tricycle graph G satisfies $\operatorname{dgon}(G) = 6$ and $\operatorname{dgon}(\Gamma(G)) = \operatorname{dgon}(\sigma_2(G)) = 5$.

4.5 A family of examples with larger gaps

In this section, we combine tricycle graphs in a certain way in order to obtain graphs G_k with $\operatorname{dgon}(G_k) = 6k$ and $\operatorname{dgon}(\sigma_2(G_k)) = \operatorname{dgon}(\Gamma(G_k)) = 5k$, which shows that the gap between $\operatorname{dgon}(\Gamma(G))$ and $\operatorname{dgon}(G)$ can be arbitrarily large. Furthermore, we show that $\operatorname{dgon}_r(\Gamma(G))$ and $\operatorname{dgon}_r(G)$ differ by at most a factor 2.

Definition 4.25. Given a (connected) simple graph H and an integer $t \geq 1$, an (H, t)-*skewered graph* is a graph G that can be obtained in the following way:

- Start with a disjoint union of graphs $G_1, \ldots, G_n$, where $n = |V(H)|$.
- For every $i \in [n]$, choose a base vertex $w_i \in V(G_i)$;
- For every edge $ij \in E(H)$, add t parallel edges between w_i and w_j, and subdivide these edges in an arbitrary way.

An example of a $(K_2, 12)$-skewered graph is given in Figure 4.3 below.

Lemma 4.26. *Let G be an (H, t)-skewered graph with $t \geq \sum_{i=1}^{|V(H)|} \operatorname{dgon}(G_i)$. Then* $\operatorname{dgon}(G) = \sum_{i=1}^{|V(H)|} \operatorname{dgon}(G_i)$.

Proof. First, we prove that $\operatorname{dgon}(G) \leq \sum_{i=1}^{|V(H)|} \operatorname{dgon}(G_i)$. For every i, choose a positive rank divisor $D_i \in \operatorname{Div}(G_i)$ of minimum degree. This defines a divisor $D \in \operatorname{Div}(G)$ with $\deg(D) = \sum_{i=1}^{|V(H)|} \operatorname{dgon}(G_i)$. We prove that D has positive rank. By Corollary 4.9, it suffices to prove that D reaches all vertices of every G_i. Let $v \in V(G_i)$, and choose an effective divisor $D_i' \in \operatorname{Div}(G_i)$ equivalent to D_i with $D_i'(v) > 0$. By Proposition 2.1, we can go from D_i to D_i' by subsequently firing an increasing sequence $U_1 \subseteq \cdots \subseteq U_k \subseteq V(G_i)$ of valid sets. Define $U_1' \subseteq \cdots \subseteq U_k' \subseteq V(G)$ by

$$U_j' := \begin{cases} U_j, & \text{if } w_i \notin U_j; \\ U_j \cup V(G_i)^c, & \text{if } w_i \in U_j. \end{cases}$$

Then, starting with D and subsequently firing the sets $U_1' \subseteq \cdots \subseteq U_k'$, we obtain an equivalent divisor $D' = D - D_i + D_i' \in \operatorname{Div}(G)$. In other words, we can play the chip-firing game on G_i while leaving the remainder of G unchanged. This shows that D reaches all vertices of every G_i, so it follows from Corollary 4.9 that $\operatorname{rank}(D) \geq 1$.

Next, we prove that $\operatorname{dgon}(G) \geq \sum_{i=1}^{|V(H)|} \operatorname{dgon}(G_i)$. Suppose, for the sake of contradiction, that $D \in \operatorname{Div}(G)$ is a positive rank w_1-reduced divisor with $\deg(D) < \sum_{i=1}^{|V(H)|} \operatorname{dgon}(G_i)$. We claim that D is w_i-reduced for all i. To that end, let $S \subseteq V(G)$ be a subset for which there is some $ij \in E(H)$ with $w_i \in S$ and $w_j \notin S$. Since there are t parallel paths in G between w_i and w_j, it follows from the max-flow min-cut theorem that $|E(S, S^c)| \geq t$. Therefore, $|E(S, S^c)| \geq t \geq \sum_{i=1}^{|V(H)|} \operatorname{dgon}(G_i) > \deg(D)$, so S cannot be fired. Thus, if $S \subseteq V(G)$ is a subset which can be fired, then $w_1 \in S$ (because D is w_1-reduced), and therefore $w_i \in S$ for all i (because H is connected). This proves our claim that D is w_i-reduced for all i.

Next, we claim that D restricts to a positive rank divisor on every G_i. Indeed, let $v \in V(G_i)$ for some i, and choose an equivalent effective divisor $D' \in \operatorname{Div}(G)$ with $D'(v) > 0$. By Proposition 2.1, we can go from D to D' by subsequently firing an increasing sequence $U_1 \subseteq \cdots \subseteq U_k$ of valid sets. Since D is w_i-reduced, we have $w_i \in U_1$, and therefore $w_i \in U_j$ for all j. Since w_i is the only vertex in G_i connected to anything outside of G_i, the firing sequence $U_1 \subseteq \cdots \subseteq U_k$ only ever sends chips out of G_i, and never into G_i. Hence it restricts to a valid firing sequence in G_i, which shows that the restricted divisor $D|_{G_i} \in \operatorname{Div}(G_i)$ reaches v. This proves our claim that D restricts to a positive rank divisor on every G_i. But now it follows that $\deg(D) \geq \sum_{i=1}^{|V(H)|} \operatorname{dgon}(G_i)$, contrary to our assumption. This is a contradiction. ■

Proof of Theorem 4.5. Let $G_1, \ldots, G_k$ be tricycle graphs, and let H be an arbitrary connected simple graph on k vertices. Choose $t \geq 6k$, and let G be an (H, t)-skewered graph obtained from the graphs $G_1, \ldots, G_k$. Then it follows from Lemma 4.26 that $\operatorname{dgon}(G) = 6k$. Furthermore, for every $s \in \mathbb{N}_1$, the subdivided graph $\sigma_s(G)$ is an (H, t)-skewered graph relative to the base graphs $\sigma_s(G_1), \ldots, \sigma_s(G_k)$, so it follows from Lemma 4.26 and Corollary 4.23 that

$$\operatorname{dgon}(\sigma_s(G)) = \sum_{i=1}^{k} \operatorname{dgon}(\sigma_s(G_i)) \geq \sum_{i=1}^{k} \operatorname{dgon}(\Gamma(G_i)) = 5k,$$

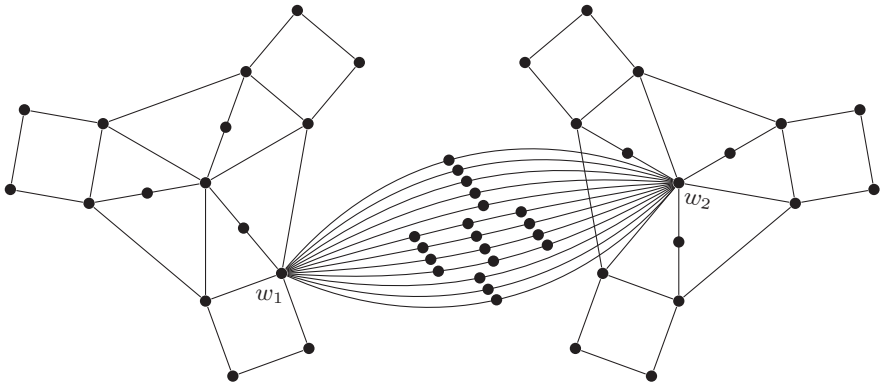


Figure 4.3: A simple, bipartite, $(K_2, 12)$-skewered tricycle graph G with $\mathrm{dgon}(G) = 12$ and $\mathrm{dgon}(\Gamma(G)) = \mathrm{dgon}(\sigma_2(G)) = 10$.

with equality if $s = 2$. Therefore $\mathrm{dgon}(\Gamma(G)) = \mathrm{dgon}(\sigma_2(G)) = 5k$.

A simple and bipartite realization can be obtained by choosing the tricycles $G_1, \ldots, G_k$ simple and bipartite (e.g. the tricycles skewered together in Figure 4.3), and choosing an appropriate subdivision in the process of Definition 4.25. ■

Theorem 4.5 shows that the discrete and metric divisorial gonality can be arbitrarily far apart. The following simple result shows that large gaps like this can only occur when the metric gonality is also large.

Proposition 4.27. *For every graph G and every integer $r \geq 1$, one has* $\mathrm{dgon}_r(G) \leq 2\,\mathrm{dgon}_r(\Gamma(G)) - r$.

Proof. Let $D_1 \in \mathrm{Div}(\Gamma(G))$ be a divisor of rank r and degree $d := \mathrm{dgon}_r(\Gamma(G))$. Choose some $E \in \mathrm{Div}^r_+(G)$, and choose a divisor $D_1' \sim D_1$ such that $D_1' \geq E$. Let $D_2 \in \mathrm{Div}(G)$ be the divisor obtained from D_1' by replacing every chip on the interior of some edge $uv \in E(G)$ by one chip on u and one chip on v. Since $D_1' \geq E$ and $\mathrm{supp}(E) \subseteq V(G)$, the divisor D_1' has at least r chips on vertices of G, so $\deg(D_2) \leq 2d - r$. By firing everything but the interior of the edge uv, we can move the newly added chips on u and v so that one of the two reaches the original position of the chip in D_1' and the other becomes superfluous. This shows that D_2 is equivalent on Γ to a divisor D_2' with $D_2' \geq D_1'$, so $\mathrm{rank}_G(D_2) = \mathrm{rank}_\Gamma(D_2) \geq r$, by [HKN13, Thm. 1.3]. ■

4.6 Computational results and open questions

Apart from the tricycle graphs, we have found a few other counterexamples, which we sketch here. First of all, the proofs from §4.4 still hold if each of the cycles C_1, C_2 and C_3 is replaced by any graph C which has two distinct vertices v^-, v^+ such that: (i) there are two edge-disjoint paths between v^- and v^+; (ii) the divisor $v^- + v^+$ has positive rank on C.

Second, we have found a number of counterexamples which we have verified computationally (using a brute force gonality algorithm), but for which we have no human-readable proof. Most of these have a structure very similar to a tricycle graph: there are 3 cycles which are connected to one another and to a central vertex in some way. A small selection of these counterexamples is given in Figure 4.4. In each of these, the optimal divisor on the 2-regular subdivision $\sigma_2(G)$ has 3 chips on the midpoints of certain edges, and 2 or 3 chips on the central vertex, and $\operatorname{dgon}(G) = \operatorname{dgon}(\sigma_2(G)) + 1$. Note that the counterexample depicted in Figure 4.4(c) is 3-regular. We have also found counterexamples where the outer ring has 5 or 7 cycles; see Figure 4.4(d). We have not found a counterexample with 9 or more cycles on the outer ring. See [DSW21] for code and additional figures.

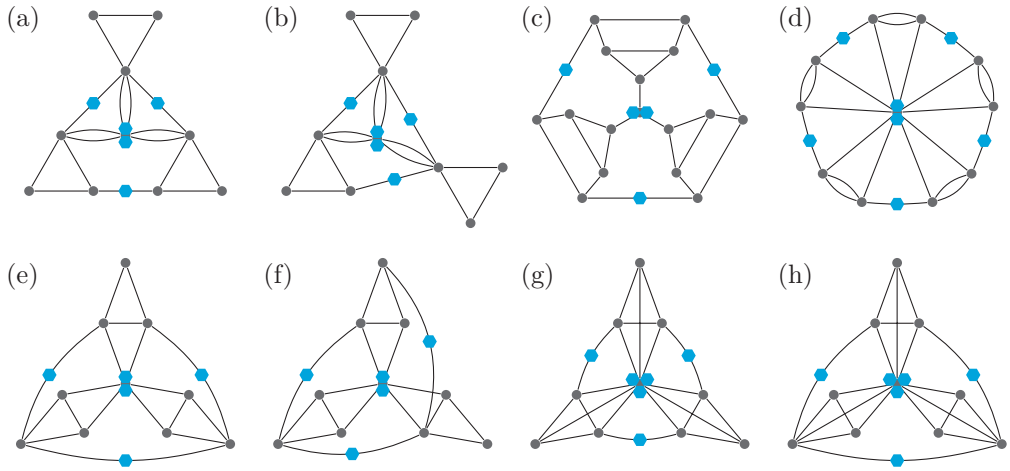


Figure 4.4: Additional counterexamples to Conjecture 4.3(a) for $k = 2$ and $r = 1$. The small cyan-coloured hexagons represent the chips of an optimal divisor on the 2-regular subdivision. In each example, the divisorial gonality of the original graph is one higher.

We have tested Conjecture 4.3(a) for $k = 2$ and $r = 1$ for all simple connected graphs on at most 10 vertices. These graphs were generated using the program `geng` from the `gtools` suite packaged with `nauty` [MP14, MP20], and tested using custom code that we wrote to compute the divisorial gonality of a graph [DSW21]. We have found that every simple connected graph with 9 or fewer vertices satisfies $\operatorname{dgon}(\sigma_2(G)) = \operatorname{dgon}(G)$, and that there are exactly 29 counterexamples with 10 vertices (and no parallel edges), including the minimal simple tricycle T_{ms} and the graphs depicted in Figure 4.4(e)–(h). All 29 minimal simple counterexamples are depicted in Figure 4.5. For code to reproduce this list, see [DSW21]. There we have also included optimized code to check whether the divisorial gonality of a given graph satisfies the Brill–Noether bound, which we have used to verify Conjecture 4.2 for all simple connected graphs with at most 13 vertices. No counterexamples were found.

We close with a few open problems.

1. As mentioned before, the Brill–Noether conjecture [Bak08, Conj. 3.9(1)] remains open.

2. What is the smallest constant c such that $\operatorname{dgon}(G) \leq c \operatorname{dgon}(\Gamma(G))$ for all graphs G? Our examples from Theorem 4.5 show that $c \geq \frac{6}{5}$, and Proposition 4.27 shows that $c \leq 2$.

3. All counterexamples presented in this chapter satisfy $\operatorname{dgon}(\sigma_2(G)) < \operatorname{dgon}(G)$. Note that this implies that $\operatorname{dgon}(\sigma_k(G)) < \operatorname{dgon}(G)$ for every even number k. Is there also a graph G such that $\operatorname{dgon}(\sigma_k(G)) < \operatorname{dgon}(G)$ for some odd number k, or a graph G such that $\operatorname{dgon}(\sigma_2(G)) = \operatorname{dgon}(G)$ but $\operatorname{dgon}(\sigma_k(G)) < \operatorname{dgon}(G)$ for some $k > 2$?

4. Is there a graph G such that $\operatorname{dgon}(\Gamma(G)) = \operatorname{dgon}(G)$, but $\operatorname{dgon}_r(\Gamma(G)) < \operatorname{dgon}_r(G)$ for some $r \geq 2$?

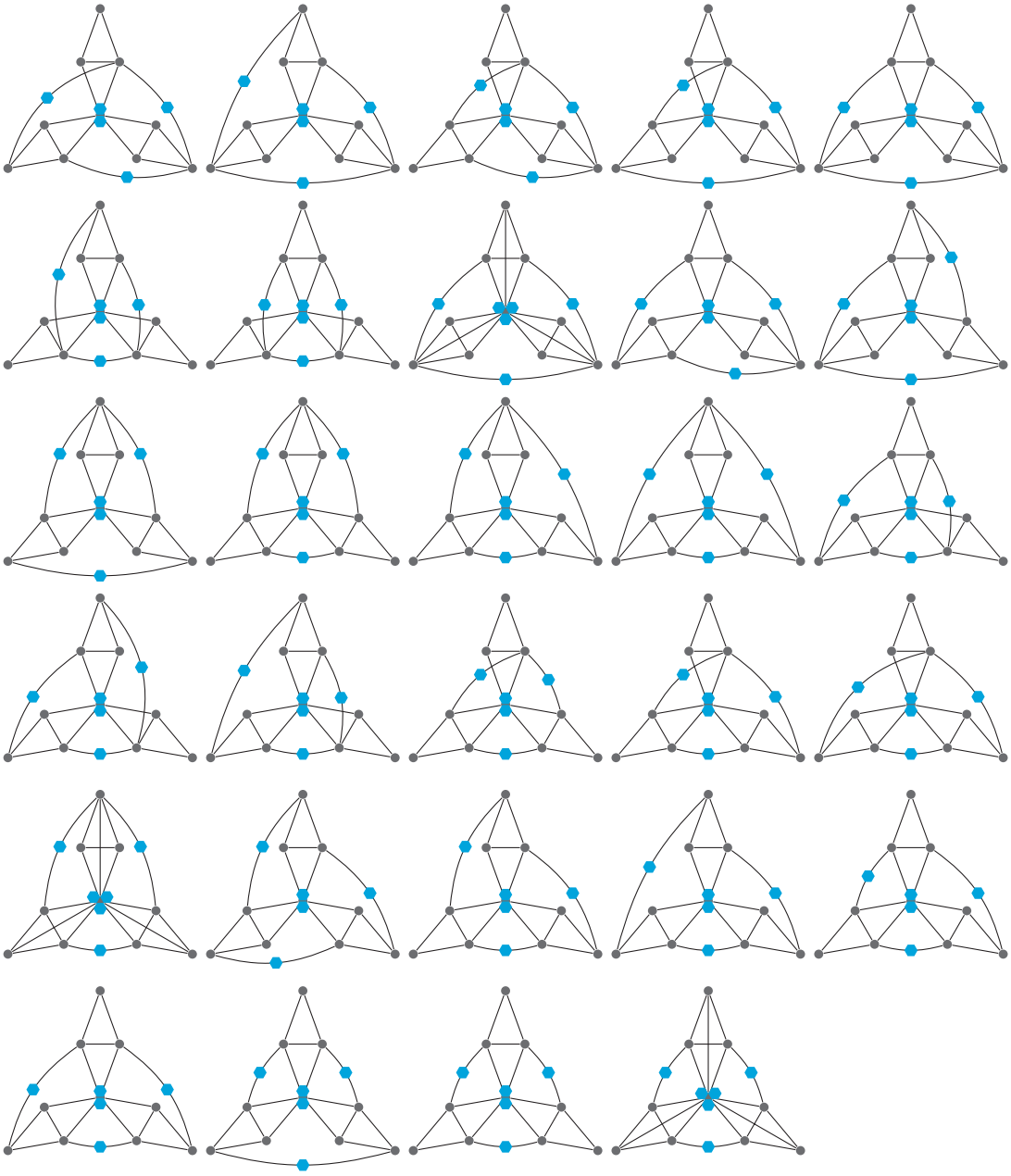


Figure 4.5: All 29 minimal simple counterexamples to Conjecture 4.3(a), with 10 vertices and no parallel edges. The small cyan-coloured hexagons represent the chips of an optimal divisor on the 2-regular subdivision. In each example, the divisorial gonality of the original graph is one higher.

Part II

The slice rank polynomial method

Chapter 5

The slice rank method

The *slice rank method* is a new technique in extremal combinatorics, developed in 2016 by Croot, Lev and Pach [CLP17], Ellenberg and Gijswijt [EG17], and Tao [Tao16]. It has been used to solve the cap set problem and make progress on several other open problems in extremal combinatorics, and will be the main focus of Part II of this dissertation. In this introductory chapter, we cover the basics of this technique, and we show how it can be applied to problems in extremal combinatorics.

Introduction

For several decades, the *cap set problem* had been one of the central open problems in extremal combinatorics. This problem asks whether or not there exists a constant $c < 3$ such that every subset $S \subseteq \mathbb{F}_3^n$ of size at least c^n contains an affine line. It is related to many other open problems in combinatorics, such as the Erdős–Turán conjecture on arithmetic progressions, the sunflower conjecture, and the computational complexity of matrix multiplication [ASU13].

In 2016, the cap set problem was solved by Ellenberg and Gijswijt [EG17], using a new technique developed earlier that year by Croot, Lev and Pach [CLP17]. Their proof was subsequently recast by Tao in terms of a new rank function for tensors, called *slice rank* [Tao16], which has led this set of techniques to be called the *slice rank method*. This will be the main topic of Part II of this dissertation.

In this chapter, we cover the definitions and basic properties of slice rank, and we discuss applications of the method to three problems in extremal combinatorics: the cap set problem, tricoloured sum-free sets, and sets without non-trivial solutions to a system of balanced linear equations. This last result will be the point of departure for the remainder of Part II.

5.1 Slice rank

Let $A_1, \dots, A_k$ be finite sets and let $\mathbb{F}$ be a field. A *hypermatrix* is a function $T : A_1 \times \cdots \times A_k \to \mathbb{F}$. A *slice* of a k-dimensional hypermatrix T is a $(k-1)$-dimensional hypermatrix T' of the form $T'(x_1, \dots, x_{k-1}) = T(x_1, \dots, x_{i-1}, a, x_i, \dots, x_{k-1})$ for some fixed $i \in [k]$ and $a \in A_i$. (For example, a slice of a matrix is a row or column.)

A hypermatrix is called *diagonal* if $A_1 = \cdots = A_k$ and $T(x_1, \ldots, x_k) = 0$ whenever $x_1, \ldots, x_k$ are not all equal.

The hypermatrix $T : A_1 \times \cdots \times A_k \to \mathbb{F}$ is said to have *slice rank one* if $T \neq 0$ and there exists $i \in [k]$ and functions $f : A_i \to \mathbb{F}$ and $g : A_1 \times \cdots \times A_{i-1} \times A_{i+1} \times \cdots \times A_k \to \mathbb{F}$ such that $T(x_1, \ldots, x_k) = f(x_i)g(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$ for all $(x_1, \ldots, x_k) \in A_1 \times \cdots \times A_k$. The *slice rank* of an arbitrary hypermatrix T, denoted $\mathrm{sr}(T)$, is the minimum integer r such that T can be written as the sum of r hypermatrices of slice rank one. (For a coordinate-free definition of slice rank, using tensors instead of hypermatrices, see [TS16].)

For $k = 2$, the slice rank is simply the matrix rank, so slice rank is a generalization of matrix rank to hypermatrices. However, it differs from the usual notion of (tensor) rank, as the only hypermatrices of tensor rank one are those of the form $T(x_1, \ldots, x_k) = f_1(x_1)f_2(x_2)\cdots f_k(x_k)$.

We proceed to prove the basic properties of the slice rank.

Proposition 5.1. *Let $T : A_1 \times \cdots \times A_k \to \mathbb{F}$ be a hypermatrix. Then:*

(a) one has $\mathrm{sr}(T) \leq \min(|A_1|, \ldots, |A_k|)$;

(b) for all subsets $B_1 \subseteq A_1, \ldots, B_k \subseteq A_k$, one has $\mathrm{sr}(T|_{B_1 \times \cdots \times B_k}) \leq \mathrm{sr}(T)$;

(c) if $T = \sum_{i=1}^{k} \sum_{j=1}^{r_i} f_{ij}(x_i)g_{ij}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$ is an optimal slice rank decomposition of T (that is, if $\mathrm{sr}(T) = \sum_{i=1}^{k} r_i$), then for each $i \in [k]$ the vectors $f_{i1}, \ldots, f_{ir_i} \in \mathbb{F}^{A_i}$ are linearly independent, and likewise for each $i \in [k]$ the slices $g_{i1}, \ldots, g_{ir_i} \in \mathbb{F}^{A_1 \times \cdots \times A_{i-1} \times A_{i+1} \times \cdots \times A_k}$ are linearly independent.

Proof. (a) For every $i \in [k]$ we have

$$T(x_1, \ldots, x_k) = \sum_{a \in A_i} \chi_a(x_i) T(x_1, \ldots, x_{i-1}, a, x_{i+1}, \ldots, x_k),$$

where $\chi_a : A_i \to \mathbb{F}$ denotes the indicator vector of $\{a\}$. This is a valid slice rank decomposition of T, so we have $\mathrm{sr}(T) \leq \min(|A_1|, \ldots, |A_k|)$.

(b) Every slice rank decomposition $T = \sum_{i=1}^{k} \sum_{j=1}^{r_i} f_{ij}(x_i)g_{ij}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$ of T restricts to a slice rank decomposition of $T|_{B_1 \times \cdots \times B_k}$, so $\mathrm{sr}(T|_{B_1 \times \cdots \times B_k}) \leq \mathrm{sr}(T)$.

(c) Suppose that $\lambda_1 f_{i1} + \cdots + \lambda_{r_i} f_{ir_i} = 0$ with $\lambda_1, \ldots, \lambda_{r_i}$ not all equal to zero. By rearranging, we may assume that $\lambda_{r_i} \neq 0$. But then we have $f_{ir_i} = -\frac{1}{\lambda_{r_i}}(\lambda_1 f_{i1} + \cdots + \lambda_{r_i - 1} f_{i,r_i-1})$, hence

$$\sum_{j=1}^{r_i} f_{ij}(x_i)g_{ij}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$$
$$= \sum_{j=1}^{r_i - 1} f_{ij}(x_i)\big(g_{ij}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k) - \tfrac{1}{\lambda_{r_i}} g_{ir_1}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)\big).$$

This reduces the number of terms in the slice rank decomposition by one, contrary to the assumption that the original decomposition was optimal. Therefore $f_{i1}, \dots, f_{ir_i}$ must be linearly independent. Analogously, $g_{i1}, \dots, g_{ir_i}$ are linearly independent. ■

In applications of the slice rank method to problems in extremal combinatorics, one must compute the slice rank of an infinite family of (ever growing) hypermatrices. This can be very challenging, as even computing the slice rank of a single hypermatrix is NP-hard [BIL+21]. However, for certain special families of hypermatrices, the slice rank is relatively easy to compute. In particular, in Proposition 5.3 below we prove a direct formula for the slice rank of a diagonal hypermatrix. For this we need the following lemma.

Lemma 5.2. *Let $\mathbb{F}$ be a field and let $V \subseteq \mathbb{F}^n$ be a linear subspace. Then there exists a vector $v \in V$ with at least $\dim(V)$ non-zero entries.*

Proof. Choose a basis $\mathcal{B} = \{b_1, \dots, b_d\}$ of V, and let M be the matrix whose rows are $b_1, \dots, b_d$. After performing Gaussian elimination, we obtain a matrix M' in reduced row echelon form which is row equivalent to M. If $b'_1, \dots, b'_d$ denote the rows of M', then $b'_1 + \cdots + b'_d$ has at least $\dim(V)$ non-zero entries. ■

Proposition 5.3 ([Tao16, Lemma 1])**.** *For all $k \geq 2$, the slice rank of every k-dimensional diagonal hypermatrix equals the number of non-zero entries.*

Proof. It is easy to see that a hypermatrix with d non-zero entries has slice rank at most d, so it suffices to show that the slice rank of a diagonal hypermatrix with d non-zero entries is at least d. We proceed by induction on k.

- For $k = 2$, the slice rank coincides with the ordinary matrix rank, and we know from linear algebra that the rank of a diagonal matrix equals the number of non-zero entries.

- Let $k \geq 3$ be given such that the slice rank of every $(k-1)$-dimensional diagonal hypermatrix is at least the number of non-zero entries. Now let $T : A^k \to \mathbb{F}$ be a k-dimensional diagonal hypermatrix. Write $A' := \{a \in A : T(a, \dots, a) \neq 0\}$, and let $T' := T|_{(A')^k}$ be the restriction of T to $(A')^k$. Choose an optimal slice rank decomposition $T' = \sum_{i=1}^{k} \sum_{j=1}^{r_i} f_{ij}(x_i) g_{ij}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_k)$ of T', so that $\mathrm{sr}(T') = \sum_{i=1}^{k} r_i$.

  Let $V := \{f_{k1}, \dots, f_{kr_k}\}^{\perp} \subseteq \mathbb{F}^{A'}$ be the space of functions $h : A' \to \mathbb{F}$ satisfying $\sum_{x \in A'} f_{kj}(x) h(x) = 0$ for all $j \in [r_k]$. By Proposition 5.1(c), we have $\dim(V) = |A'| - r_k$, so by Lemma 5.2 we may choose some $h \in V$ with at least $|A'| - r_k$ non-zero entries. Define $T'' : (A')^{k-1} \to \mathbb{F}$ by

$$T''(x_1, \dots, x_{k-1}) := \sum_{x_k \in A'} T'(x_1, \dots, x_k) h(x_k).$$

Then T'' is a $(k-1)$-dimensional diagonal hypermatrix with at least $|A'|-r_k$ nonzero entries, so it follows from the induction hypothesis that $\mathrm{sr}(T'') \geq |A'| - r_k$. On the other hand, since $h \in \{f_{k1}, \dots, f_{kr_k}\}^\perp$, we have

$$\sum_{x_k \in A'} \sum_{j=1}^{r_k} f_{kj}(x_k) g_{kj}(x_1, \dots, x_{k-1}) h(x_k) = \sum_{j=1}^{r_k} g_{kj}(x_1, \dots, x_{k-1}) \sum_{x_k \in A'} f_{kj}(x_k) h(x_k) = 0,$$

so plugging in the slice rank decomposition of T' gives

$$\begin{aligned} T''(x_1, \dots, x_{k-1}) &= \sum_{x_k \in A'} \sum_{i=1}^{k} \sum_{j=1}^{r_i} f_{ij}(x_i) g_{ij}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_k) h(x_k) \\ &= \sum_{i=1}^{k-1} \sum_{x_k \in A'} \sum_{j=1}^{r_i} f_{ij}(x_i) g_{ij}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_k) h(x_k) \\ &= \sum_{i=1}^{k-1} \sum_{j=1}^{r_i} f_{ij}(x_i) \cdot \left(\sum_{x_k \in A'} g_{ij}(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_k) h(x_k) \right). \end{aligned}$$

This gives a valid slice rank decomposition of T'' of size $r_1 + \dots + r_{k-1}$, so we have $\mathrm{sr}(T'') \leq r_1 + \dots + r_{k-1}$. It follows that $|A'| \leq \mathrm{sr}(T'') + r_k \leq r_1 + \dots + r_k = \mathrm{sr}(T')$, so by Proposition 5.1(b) we have $\mathrm{sr}(T) \geq \mathrm{sr}(T') \geq |A'|$. ■

In all applications of the slice rank method in this dissertation, the hypermatrix will be diagonal, so Proposition 5.3 is all we need. However, we point out that slice rank formulas are known for certain other classes of hypermatrices as well; see for instance [TS16, Sau22].

5.2 Monomials of small degree and the Croot–Lev–Pach lemma

In a typical application of the slice rank method, the hypermatrix under consideration is defined by a polynomial which encodes the combinatorial structure of the problem. If this polynomial has sufficiently low degree, then a clever expansion of this polynomial and a large deviations bound show that the hypermatrix has exponentially small size. In this section, we make these statements precise.

Let $\mathbb{F}$ be a field, and let $\mathbb{F}[x_1, \dots, x_n]$ be the polynomial ring in n variables over $\mathbb{F}$. For $\boldsymbol{\alpha} \in \mathbb{N}_0^n$, write $\boldsymbol{x}^{\boldsymbol{\alpha}} := x_1^{\alpha_1} \cdots x_n^{\alpha_n} \in \mathbb{F}[x_1, \dots, x_n]$ and $|\boldsymbol{\alpha}| = \alpha_1 + \dots + \alpha_n$. Moreover, for $q \in \mathbb{N}_1$ and $d \in \mathbb{R}_{\geq 0}$, write $M_{q,n,d} := \{\boldsymbol{\alpha} \in \{0, 1, \dots, q-1\}^n : |\boldsymbol{\alpha}| \leq d\}$ and $m_{q,n,d} := |M_{q,n,d}|$. Then $m_{q,n,d}$ is equal to the number of monomials in $\mathbb{F}[x_1, \dots, x_n]$ whose degree in every variable (separately) is at most $q-1$ and whose total degree is at most d.[1]

The following lemma shows that a hypermatrix defined by a polynomial of small total degree has relatively small slice rank. It is named after Croot, Lev and Pach,

[1] For technical reasons, it will be convenient to also allow non-integer values of d, even though the total degree of a monomial is always an integer.

whose paper [CLP17] led to the solution of the cap set problem [EG17] and the development of the slice rank method [Tao16]. The $k = 2$ case of the lemma is implicit in [CLP17].

Lemma 5.4 (Generalized Croot–Lev–Pach lemma)**.** *Let q be a prime power, let $n \in \mathbb{N}_1$, and let $A_1, \dots, A_k \subseteq \mathbb{F}_q^n$. Furthermore, let $T : A_1 \times \cdots \times A_k \to \mathbb{F}_q$ be a hypermatrix given by*

$$T(\boldsymbol{x_1}, \dots, \boldsymbol{x_k}) = f(x_{11}, \dots, x_{1n}, \dots, x_{k1}, \dots, x_{kn})$$

for some polynomial $f \in \mathbb{F}_q[x_{11}, \dots, x_{1n}, \dots, x_{k1}, \dots, x_{kn}]$ of total degree at most d. Then $\operatorname{sr}(T) \le k \cdot m_{q,n,\frac{d}{k}}$.

Proof. Since $x^q = x$ for all $x \in \mathbb{F}_q$, we may assume without loss of generality that f has degree at most $q-1$ in every variable x_{ij} (separately), and total degree at most d.

For $i \in [k]$ and $\boldsymbol{\alpha_i} \in \{0, 1, \dots, q-1\}^n$, write $\boldsymbol{x_i^{\alpha_i}} = x_{i1}^{\alpha_{i1}} \cdots x_{in}^{\alpha_{in}}$ and $|\boldsymbol{\alpha_i}| = \alpha_{i1} + \cdots + \alpha_{in}$. Furthermore, write $\boldsymbol{M_{q,n,k,d}} = \{(\boldsymbol{\alpha_1}, \dots, \boldsymbol{\alpha_k}) \in (\{0, 1, \dots, q-1\}^n)^k : |\boldsymbol{\alpha_1}| + \cdots + |\boldsymbol{\alpha_k}| \le d\}$. Then we may write f as

$$f(x_{11}, \dots, x_{1n}, \dots, x_{k1}, \dots, x_{kn}) = \sum_{(\boldsymbol{\alpha_1}, \dots, \boldsymbol{\alpha_k}) \in \boldsymbol{M_{q,n,k,d}}} \xi_{(\boldsymbol{\alpha_1}, \dots, \boldsymbol{\alpha_k})} \cdot \boldsymbol{x_1^{\alpha_1}} \cdots \boldsymbol{x_k^{\alpha_k}},$$

where $\xi_{(\boldsymbol{\alpha_1}, \dots, \boldsymbol{\alpha_k})} \in \mathbb{F}_q$ for all $(\boldsymbol{\alpha_1}, \dots, \boldsymbol{\alpha_k}) \in \boldsymbol{M_{q,n,k,d}}$.

By the pigeonhole principle, for every $(\boldsymbol{\alpha_1}, \dots, \boldsymbol{\alpha_k}) \in \boldsymbol{M_{q,n,k,d}}$ there is some $i \in [k]$ such that $|\boldsymbol{\alpha_i}| \le \frac{d}{k}$. Choosing one such i for every $(\boldsymbol{\alpha_1}, \dots, \boldsymbol{\alpha_k}) \in \boldsymbol{M_{q,n,k,d}}$, we obtain a partition $\boldsymbol{M_{q,n,k,d}} = \mathcal{M}_1 \cup \cdots \cup \mathcal{M}_k$ such that for all $(\boldsymbol{\alpha_1}, \dots, \boldsymbol{\alpha_k}) \in \mathcal{M}_i$ we have $|\boldsymbol{\alpha_i}| \le \frac{d}{k}$.

Now we can write

$$\begin{aligned} T(\boldsymbol{x_1}, \dots, \boldsymbol{x_k}) &= \sum_{(\boldsymbol{\alpha_1}, \dots, \boldsymbol{\alpha_k}) \in \boldsymbol{M_{q,n,k,d}}} \xi_{(\boldsymbol{\alpha_1}, \dots, \boldsymbol{\alpha_k})} \cdot \boldsymbol{x_1^{\alpha_1}} \cdots \boldsymbol{x_k^{\alpha_k}} \\ &= \sum_{i=1}^{k} \sum_{(\boldsymbol{\alpha_1}, \dots, \boldsymbol{\alpha_k}) \in \mathcal{M}_i} \xi_{(\boldsymbol{\alpha_1}, \dots, \boldsymbol{\alpha_k})} \cdot \boldsymbol{x_1^{\alpha_1}} \cdots \boldsymbol{x_k^{\alpha_k}} \\ &= \sum_{i=1}^{k} \sum_{\boldsymbol{\alpha} \in M_{q,n,d/k}} \boldsymbol{x_i^{\alpha}} \left(\sum_{\substack{(\boldsymbol{\alpha_1}, \dots, \boldsymbol{\alpha_k}) \in \mathcal{M}_i \\ \boldsymbol{\alpha_i} = \boldsymbol{\alpha}}} \xi_{(\boldsymbol{\alpha_1}, \dots, \boldsymbol{\alpha_k})} \cdot \boldsymbol{x_1^{\alpha_1}} \cdots \boldsymbol{x_{i-1}^{\alpha_{i-1}}} \boldsymbol{x_{i+1}^{\alpha_{i+1}}} \cdots \boldsymbol{x_k^{\alpha_k}} \right). \end{aligned}$$

This is a valid slice rank decomposition of size $k \cdot m_{q,n,\frac{d}{k}}$, so we have $\operatorname{sr}(T) \le k \cdot m_{q,n,\frac{d}{k}}$. ∎

The remainder of this section is devoted to upper bounding $m_{q,n,d}$. For $q \in \mathbb{N}_1$ and $\lambda \in \mathbb{R}_{\ge 0}$, define

$$J(q, \lambda) := \inf_{t > 0} \frac{1 + t + t^2 + \cdots + t^{q-1}}{t^\lambda}.$$

The following estimate shows that $m_{q,n,\lambda n}$ is small when $\lambda < \frac{q-1}{2}$ is fixed and $n \to \infty$. We give a short, self-contained proof, but we note that this follows from a more general principle in probability theory; see Remark 5.6.

Lemma 5.5. *For all integers $n \geq 1$ and $q \geq 2$ and all $\lambda \in \mathbb{R}$ with $0 \leq \lambda < \frac{q-1}{2}$, one has*

$$1 < J(q,\lambda) < q \qquad \text{and} \qquad m_{q,n,\lambda n} \leq J(q,\lambda)^n.$$

Proof. Define $f : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ by $f(t) := \frac{1+t+t^2+\cdots+t^{q-1}}{t^\lambda}$. For all $t > 1$ we have

$$f(t) = \frac{1+t+t^2+\cdots+t^{q-1}}{t^\lambda} > \frac{1+t+t^2+\cdots+t^{q-1}}{t^{\frac{q-1}{2}}} \geq q = f(1),$$

where the first inequality holds because $t > 1$ and $\lambda < \frac{q-1}{2}$ and the second inequality follows from the AM–GM inequality (applied to the sequence $1, t, t^2, \ldots, t^{q-1}$). Therefore the infimum can be taken over $(0,1]$ instead of $(0,\infty)$; that is, we have $J(q,\lambda) = \inf_{t\in(0,1]} f(t)$.

We have $f'(t) = \sum_{i=0}^{q-1}(i-\lambda)t^{i-\lambda-1}$, so in particular $f'(1) = \sum_{i=0}^{q-1}(i-\lambda) = \frac{q(q-1-2\lambda)}{2} > 0$, because $q > 0$ and $q - 1 - 2\lambda > 0$, by assumption. It follows that f is strictly increasing in some neighbourhood $(1-\varepsilon, 1+\varepsilon)$ of 1, so we have $J(q,\lambda) \leq f(1-\frac{1}{2}\varepsilon) < f(1) = q$. Furthermore, since f is continuous and $\lim_{t\to 0} f(t) = +\infty$, the function f has a minimum on $(0,1]$, so there is some $t_0 \in (0,1]$ such that $J(q,\lambda) = f(t_0)$. Therefore,

$$J(q,\lambda) = \frac{1+t_0+t_0^2+\cdots+t_0^{q-1}}{t_0^\lambda} \geq 1+t_0+t_0^2+\cdots+t_0^{q-1} > 1,$$

where the first inequality holds because $0 < t_0 \leq 1$ and $\lambda \geq 0$ and the second inequality holds because $q \geq 2$ and $t_0 > 0$. This shows that $1 < J(q,\lambda) < q$.

In order to prove that $m_{q,n,\lambda n} \leq J(q,\lambda)^n$, write $(1+t+\cdots+t^{q-1})^n = \sum_{i=0}^{(q-1)n} b_i t^i$, where b_i is the number of $\boldsymbol{\alpha} \in \{0,1,\ldots,q-1\}^n$ with $|\boldsymbol{\alpha}| = i$. Then for all $t \in (0,1]$ we have

$$\begin{aligned}
m_{q,n,\lambda n} &= \sum_{i=0}^{\lfloor \lambda n \rfloor} b_i \\
&\leq \sum_{i=0}^{\lfloor \lambda n \rfloor} b_i t^{i-\lambda n} && \text{(because } t^{i-\lambda n} \geq 1 \text{ whenever } t \in (0,1] \text{ and } i \leq \lambda n\text{)} \\
&\leq \sum_{i=0}^{(q-1)n} b_i t^{i-\lambda n} && \text{(because } b_i t^{i-\lambda n} \geq 0 \text{ for all } i > \lfloor \lambda n \rfloor\text{)} \\
&= t^{-\lambda n} \sum_{i=0}^{(q-1)n} b_i t^i \;=\; f(t)^n.
\end{aligned}$$

Minimizing over $t \in (0,1]$ shows that $m_{q,n,\lambda n} \leq J(q,\lambda)^n$. ∎

Remark 5.6. We note that Lemma 5.5 also follows from a more general principle in probability theory, which is perhaps a bit more insightful than the 'ad hoc' proof given above. The problem of bounding $m_{q,n,\lambda n}$ can be modelled as the following

probability experiment. Let $X_1, \dots, X_n$ be i.i.d. random variables, drawn from the uniform distribution on $\{0, 1, \dots, q-1\}$. Then $m_{q,n,\lambda n} = q^n \cdot \Pr[X_1 + \dots + X_n \leq \lambda n]$, so to bound $m_{q,n,\lambda n}$ we need to bound the probability of the event $\{X_1 + \dots + X_n \leq \lambda n\}$.

By the central limit theorem, the normalized partial sums $\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)$ tend to a standard normal distribution. This tells us that events of the form $\{X_1 + \dots + X_n \leq \mu n - c\sqrt{n}\}$ (with $c > 0$ fixed) occur with constant probability as $n \to \infty$, whereas events of the form $\{X_1 + \dots + X_n \leq \mu n - cn\}$ (with $c > 0$ fixed) become exceedingly rare. Since $\lambda < \frac{q-1}{2} = \mathbb{E}[X_1]$, we are dealing with a problem of the latter form. To bound these so-called *tail probabilities*, we draw on the *theory of large deviations*. For this particular problem, the rate of convergence is governed by the following theorem, known as *Cramér's theorem*.

Theorem 5.7 (Cramér, Chernoff)**.** *Let $\{X_n\}_{n \in \mathbb{N}}$ be a sequence of i.i.d. real random variables with well-defined (finite) expectation $\mu := \mathbb{E}[X_1]$. Define $\Lambda^* : \mathbb{R} \to [0, +\infty]$ by*

$$\Lambda^*(x) := \sup_{t \in \mathbb{R}} \big(tx - \log \mathbb{E}[e^{tX_1}] \big),$$

with the convention that $\log(+\infty) = +\infty$. Then for all $x < \mu$ one has

$$\Pr[X_1 + \dots + X_n \leq xn] \leq e^{-\Lambda^*(x)n} \qquad \text{for all } n \in \mathbb{N}_1;$$

$$\lim_{n \to \infty} \frac{1}{n} \log \Pr[X_1 + \dots + X_n \leq xn] = -\Lambda^*(x);$$

with the convention that $e^{-\infty} = 0$ and $\log(0) = -\infty$.

A proof of Theorem 5.7 can be found in some of the more comprehensive textbooks on general probability theory (e.g. [Bau96, Kle08, Kal21]) and in textbooks specializing in large deviations theory (e.g. [Var84, DS89, DZ98, Hol00, RS15]).[2]

In the setting of Remark 5.6, a straightforward computation shows that $e^{-\Lambda^*(\lambda)n} = \frac{1}{q^n} J(q, \lambda)^n$, so Cramér's theorem gives the same upper bound as Lemma 5.5. Therefore the limit in Cramér's theorem shows that the upper bound $m_{q,n,\lambda n} \leq J(q, \lambda)^n$ is asymptotically optimal.

5.3 The slice rank method: three examples

We now have all the ingredients to apply the slice rank method to problems in extremal combinatorics. In this section, we cover three example applications: the cap set problem, tricoloured sum-free sets, and sets without non-trivial solutions to a system of balanced linear equations.

The cap set problem

For $n \in \mathbb{N}_1$, let $a(n)$ denote the largest size of a subset $S \subseteq \mathbb{F}_3^n$ which does not contain an affine line. The asymptotic behaviour of $a(n)$ as $n \to \infty$ has been subject of study

[2] The inequality $\Pr[X_1 + \dots + X_n \leq xn] \leq e^{-\Lambda^*(x)n}$ is not always stated as part of the theorem, but it is usually contained in the proof.

for several decades. In 1982, Brown and Buhler [BB82] showed that $a(n) \in o(3^n)$; that is, $\lim_{n\to\infty} \frac{a(n)}{3^n} = 0$. A few years later, Frankl, Graham and Rödl [FGR87] gave a shorter proof of that same fact, and asked whether there exists a constant $c < 3$ such that $a(n) \in O(c^n)$. This problem became known as the *cap set problem*.

Despite substantial interest in this problem, very few improvements were made on the upper bounds for $a(n)$. In 1995, Meshulam [Mes95] showed that $a(n) \leq 2 \cdot \frac{3^n}{n}$, and only in 2012 this was improved to $a(n) \in O(\frac{3^n}{n^{1+\varepsilon}})$ for some constant $\varepsilon > 0$ by Bateman and Katz [BK12]. The Fourier-analytic techniques employed by Meshulam and Bateman and Katz appeared to have reached their limit, and it was suspected that further progress might be possible using the polynomial method [Tao10], but for a long time it was unclear how to do so.

All of this suddenly changed in 2016, when Croot, Lev and Pach published a preprint containing a new application of the polynomial method to the closely related problem of avoiding 3-term arithmetic progressions in $(\mathbb{Z}/4\mathbb{Z})^n$. Soon after, it was realized by Ellenberg and Gijswijt (independently) that the method of Croot, Lev and Pach could be modified to solve the cap set problem, and they proved the following more general result.

Theorem 5.8 (Ellenberg–Gijswijt [EG17])**.** *Let $p \geq 3$ be prime, and let $c_p := J(p, \frac{p-1}{3}) < p$. Then for every $n \in \mathbb{N}$ and every subset $S \subseteq \mathbb{F}_p^n$ without non-trivial 3-term arithmetic progressions, one has $|S| \leq c_p^n$.*

Over $\mathbb{F}_3$, the non-trivial 3-term arithmetic progressions are precisely the affine lines, so the $p = 3$ case of Theorem 5.8 provides an affirmative answer to the cap set problem.

The proof was later recast by Tao in terms of slice rank [Tao16], and this has since become the dominant terminology. We follow Tao's proof, for which we already set up all the necessary prerequisites in the previous sections.

Proof of Theorem 5.8. Let $S \subseteq \mathbb{F}_p^n$ be a set without non-trivial 3-term arithmetic progressions. Define $T : S \times S \times S \to \mathbb{F}_p$ by

$$T(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) := \prod_{i=1}^{n} (1 - (x_i - 2y_i + z_i)^{p-1}).$$

By Fermat's little theorem, we have $(x_i - 2y_i + z_i)^{p-1} = 1$ if and only if $x_i - 2y_i + z_i \neq 0$, hence

$$T(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \begin{cases} 0, & \text{if } x_i - 2y_i + z_i \neq 0 \text{ for some } i; \\ 1, & \text{if } \boldsymbol{x} - 2\boldsymbol{y} + \boldsymbol{z} = 0. \end{cases}$$

It follows that $T(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \neq 0$ if and only if $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$ form a 3-term arithmetic progression. By assumption, the only 3-term arithmetic progressions in S are the trivial ones, so we have $T(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \neq 0$ if and only if $\boldsymbol{x} = \boldsymbol{y} = \boldsymbol{z}$. This shows that T is a diagonal hypermatrix with non-zero entries on the diagonal. Hence it follows from Proposition 5.3 that $\operatorname{sr}(T) = |S|$. On the other hand, by Lemma 5.4 and Lemma 5.5, we have $\operatorname{sr}(T) \leq 3 \cdot m_{p,n,\frac{(p-1)n}{3}} \leq 3 \cdot J(p, \frac{p-1}{3})^n = 3c_p^n$. This shows that $|S| \leq 3c_p^n$.

To get rid of the additional factor 3, we use the following trick (known as the 'power trick'). Let $\ell \in \mathbb{N}_1$, and consider the set $S^\ell \subseteq (\mathbb{F}_p^n)^\ell \cong \mathbb{F}_p^{n\ell}$. If $(\boldsymbol{x_1}, \dots, \boldsymbol{x_\ell}) - 2(\boldsymbol{y_1}, \dots, \boldsymbol{y_\ell}) + (\boldsymbol{z_1}, \dots, \boldsymbol{z_\ell}) = (0, \dots, 0)$, then we must have $\boldsymbol{x_i} = \boldsymbol{y_i} = \boldsymbol{z_i}$ for all $i \in [\ell]$, so S^ℓ does not contain non-trivial 3-term arithmetic progressions. Hence, by the preceding result, we have $|S|^\ell = |S^\ell| \leq 3c_p^{n\ell}$, and therefore $|S| \leq 3^{1/\ell} \cdot c_p^n$. Letting $\ell \to \infty$, we conclude that $|S| \leq c_p^n$. ∎

It is not known whether or not the constant $c_p = J(p, \frac{p-1}{3})$ in Theorem 5.8 is optimal. Currently, the best known upper bound on the size of a set $S \subseteq \mathbb{F}_p^n$ without 3-term arithmetic progressions is $O(\frac{1}{\sqrt{n}} \cdot J(p, \frac{p-1}{3})^n)$, due to Jiang [Jia21]. In particular, for the cap set problem ($p = 3$) we have

$$c_3 = J\left(3, \tfrac{2}{3}\right) = \frac{3}{8}\sqrt[3]{207 + 33\sqrt{33}} \approx 2.755105,$$

so the best known upper bound on $a(n)$ is $O(\frac{2.755105^n}{n})$. This is still far away from the best known lower bound, which is $\Omega(2.218021^n)$, due to Tyrrell [Tyr22]. For lower bounds for other primes ($p \geq 5$), see e.g. [Ede04, EP20, EL23].

Tricoloured sum-free sets

After the cap set problem was solved, it was quickly realized that the same technique could also be used to prove an asymmetric (or 'tricoloured') version of Theorem 5.8.

Let G be an abelian group. A sequence $\{(x_i, y_i, z_i)\}_{i=1}^L$ in G^3 is called a *tricoloured sum-free set in* G if for all $i, j, k \in [L]$ one has $x_i + y_j + z_k = 0$ if and only if $i = j = k$.[3] Observe that the definition implies that $|\{x_1, \dots, x_L\}| = |\{y_1, \dots, y_L\}| = |\{z_1, \dots, z_L\}| = L$; that is, in a tricoloured sum-free set there can be no repetitions in each of the coordinates (separately).

A straightforward application of the slice rank method gives the following upper bound on the size of tricoloured sum-free sets in $\mathbb{F}_p^n$:

Theorem 5.9 (Blasiak–Church–Cohn–Grochow–Naslund–Sawin–Umans [BCC+17])**.** *Let p be prime, and let $c_p := J(p, \frac{p-1}{3}) < p$. Then for every tricoloured sum-free set $\{(\boldsymbol{x_i}, \boldsymbol{y_i}, \boldsymbol{z_i})\}_{i=1}^L$ in $\mathbb{F}_p^n$ one has $L \leq c_p^n$.*

Proof. Let $\{(\boldsymbol{x_i}, \boldsymbol{y_i}, \boldsymbol{z_i})\}_{i=1}^L$ be a tricoloured sum-free set in $\mathbb{F}_p^n$. Define $T : [L] \times [L] \times [L] \to \mathbb{F}_p$ by

$$\begin{aligned} T(i,j,k) &:= \prod_{\ell=1}^{n} (1 - (x_{i\ell} + y_{j\ell} + z_{k\ell})^{p-1}) \\ &= \begin{cases} 1, & \text{if } \boldsymbol{x_i} + \boldsymbol{y_j} + \boldsymbol{z_k} = 0; \\ 0, & \text{otherwise;} \end{cases} \end{aligned}$$

[3] Sometimes tricoloured sum-free sets are defined more generally by the property that $x_i + y_j + z_k = t$ if and only if $i = j = k$, where $t \in G$ is fixed. However, if $\{(x_i, y_i, z_i)\}_{i=1}^L$ is a generalized tricoloured sum-free set, then $\{(x_i, y_i, z_i - t)\}_{i=1}^L$ is a tricoloured sum-free set, so there is no loss in generality by restricting our attention to the $t = 0$ case.

$$= \begin{cases} 1, & \text{if } i = j = k; \\ 0, & \text{otherwise.} \end{cases}$$

By Proposition 5.3, we have $\mathrm{sr}(T) = L$. On the other hand, by Lemma 5.4 and Lemma 5.5, we have $\mathrm{sr}(T) \leq 3 \cdot m_{p,n,\frac{(p-1)n}{3}} \leq 3 \cdot J(p, \frac{p-1}{3})^n = 3c_p^n$. This shows that $|S| \leq 3c_p^n$.

To get rid of the additional factor 3, we repeat the 'power trick' from the proof of Theorem 5.8. If $\{(\boldsymbol{x_i}, \boldsymbol{y_i}, \boldsymbol{z_i})\}_{i=1}^L$ is a tricoloured sum-free set in $\mathbb{F}_p^n$ and if $\ell \in \mathbb{N}$, then $\{((\boldsymbol{x_{i_1}}, \dots, \boldsymbol{x_{i_\ell}}), (\boldsymbol{y_{i_1}}, \dots, \boldsymbol{y_{i_\ell}}), (\boldsymbol{z_{i_1}}, \dots, \boldsymbol{z_{i_\ell}}))\}_{(i_1,\dots,i_\ell)\in[L]^\ell}$ is a tricoloured sum-free set in $(\mathbb{F}_p^n)^\ell \cong \mathbb{F}_p^{n\ell}$. Therefore we have $L^\ell \leq 3 \cdot c_p^{n\ell}$, hence $L \leq 3^{1/\ell} \cdot c_p^n$. Letting $\ell \to \infty$, we conclude that $L \leq c_p^n$. ∎

In [BCC+17, Thm. A], Blasiak et al. also gave an extension of Theorem 5.9 to arbitrary abelian groups of bounded exponent. A further extension to non-abelian finite groups was given by Sawin [Saw18].

Since $\mathbb{F}_{p^s}^n \cong \mathbb{F}_p^{sn}$ as groups, the following corollary is immediate.

Corollary 5.10. *Let $q = p^s$ be a prime power, and let $c_q := J(p, \frac{p-1}{3})^s < q$. Then for every tricoloured sum-free set $\{(\boldsymbol{x_i}, \boldsymbol{y_i}, \boldsymbol{z_i})\}_{i=1}^L$ in $\mathbb{F}_q^n$ one has $L \leq c_q^n$.*

Tricoloured sum-free sets are the asymmetric (or 'tricoloured') equivalent of cap sets, in the following sense. If $S \subseteq \mathbb{F}_p^n$ is a set without non-trivial 3-term arithmetic progressions, and if $S = \{\boldsymbol{x_i}\}_{i=1}^L$ is an enumeration of the elements of S (where $L = |S|$), then $\{(\boldsymbol{x_i}, -2\boldsymbol{x_i}, \boldsymbol{x_i})\}_{i=1}^L$ is a tricoloured sum-free set. In this setting, the proofs from Theorem 5.8 and Theorem 5.9 use the same hypermatrix and result in the same upper bound. Hence Theorem 5.9 can be seen as a generalization of Theorem 5.8 where the three variables can be taken from different sets.

Contrary to the cap set problem, the bound from Theorem 5.9 is known to be tight up to a subexponential factor. In a series of papers, Kleinberg, Speyer and Sawin [KSS18], Pebody [Peb18] and (independently) Norin [Nor19] proved that for every $\varepsilon > 0$, there is a tricoloured sum-free set of size $(c_p - \varepsilon)^n$ in $\mathbb{F}_p^n$ for large enough n. This shows that no exponential improvement of the cap set bound from Theorem 5.8 is possible without somehow taking the symmetry of that problem into account.

Remark 5.11. In the next chapter, it will be convenient to replace the bound $L \leq c_q^n$ from Corollary 5.10 by a strict inequality, $L < c_q^n$. We can do this because $J(p, \frac{p-1}{3})$ is not the n-th root of an integer, but this is not completely trivial. We sketch a proof of this fact here, assuming familiarity with a bit of algebra (see e.g. [Lan02, §IV.2]).

Write $F(t) = (1 + t + \cdots + t^{p-1})t^{-\frac{p-1}{3}}$. Recall from the proof of Lemma 5.5 that F attains a minimum on $(0, 1)$. Write $t_0 = \arg\min_{t\in(0,1)} F(t)$, so that $J(p, \frac{p-1}{3}) = F(t_0)$.

First we prove that t_0 is algebraic, but not an algebraic integer (i.e. it is not the zero of a *monic* polynomial in $\mathbb{Z}[X]$). A direct computation shows that $F'(t) = \frac{1}{3} \cdot f(t) \cdot t^{-\frac{p-1}{3}-1}$, where $f(t) = (-p+1) + (-p+4)t + \cdots + (2p-2)t^{p-1}$. Since $F'(t_0) = 0$ and $t_0 \neq 0$, it follows that $f(t_0) = 0$, which shows that t_0 is algebraic.

For $p \in \{2, 3\}$ it is easy to verify that t_0 is not an algebraic integer. Assume henceforth that $p \geq 5$. Let $c = c(f)$ and $\mathrm{pp}(f)$ be the *content* and *primitive part* of f,

respectively (i.e. c is the gcd of the coefficients of f, and $\mathrm{pp}(f) = \frac{1}{c}f$). Since successive coefficients of f differ by 3, we have either $c = 1$ (if $p \not\equiv 1 \mod 3$) or $c = 3$ (if $p \equiv 1 \mod 3$), so either $\mathrm{pp}(f) = f$ or $\mathrm{pp}(f) = \frac{1}{3}f$. In either case, $\mathrm{pp}(f)$ is not monic.

Now suppose, for the sake of contradiction, that t_0 is an algebraic integer. Then the minimal polynomial $g \in \mathbb{Z}[t]$ of t_0 is monic and has integer coefficients, so it is different from $\mathrm{pp}(f)$. Therefore we may write $\mathrm{pp}(f) = gh$ for some non-constant $h \in \mathbb{Z}[t]$. Let $\bar{f}, \bar{g}, \bar{h} \in \mathbb{F}_p[t]$ be the reductions of f, g, h modulo p. A direct computation shows that $\bar{f} = -2 \cdot (t-1)^{p-2}(t - \frac{p-1}{2})$, so we have $\bar{g} = (t-1)^k(t-\frac{p-1}{2})^\ell$ and $\bar{h} = \frac{-2}{c} \cdot (t-1)^{p-2-k}(t-\frac{p-1}{2})^{1-\ell}$ for some $k \in \{0, \dots, p-2\}$ and $\ell \in \{0,1\}$. Since $f(1) = \frac{p(p-1)}{2}$ is divisible by p but not by p^2, at most one of $g(1)$ and $h(1)$ can be divisible by p, so at most one of $\bar{g}$ and $\bar{h}$ vanishes on 1. Since $\deg(g), \deg(h) \geq 1$, we must have $(k,\ell) = (p-2, 0)$ or $(k,\ell) = (0,1)$.

- If $(k,\ell) = (0,1)$, then $\deg(g) = 1$, so t_0 is a rational number. Since we assumed that t_0 is an algebraic integer, we must have $t_0 \in \mathbb{Z}$. This is a contradiction, because $0 < t_0 < 1$.

- If $(k,\ell) = (p-2, 0)$, then $\deg(h) = 1$. Write $h(t) = at - b$. By primitivity of $\mathrm{pp}(f)$, we must have $\gcd(a,b) = 1$. Since g is monic, the leading coefficient of h is equal to the leading coefficient of $\mathrm{pp}(f)$, so $a = \frac{2p-2}{c}$. Furthermore, since $\frac{1-p}{c} = \frac{1}{c}f(0) = g(0)h(0)$, we have $b = h(0) \mid \frac{1-p}{c} \mid a$. But since $\gcd(a,b) = 1$, it follows that $b = \pm 1$. Finally, since $g(1)h(1) = \frac{1}{c}f(1) = \frac{p(p-1)}{2c}$, we have $a - b = h(1) \mid \frac{p(p-1)}{2c}$. But $a - b = \frac{2p-2}{c} \pm 1$ is not divisible by p (for $c = 1$ we have $a - b \not\equiv 0 \mod p$ since we assumed $p \geq 5$, and for $c = 3$ we have $0 < a - b < p$), so in fact we have $a - b = \frac{2p-2}{c} \pm 1 \mid \frac{p-1}{c}$, which is absurd.

Either way, we reach a contradiction, so we conclude that t_0 is not an algebraic integer. (With a bit more work, one can also show that $\frac{1}{c}f$ is in fact the minimal polynomial of t_0, but we don't need that here.)

Finally, to show that $J(p, \frac{p-1}{3})$ is not the n-th root of an integer, suppose that $J(p, \frac{p-1}{3})^n = b$ for some $b \in \mathbb{Z}$. Then we have $F(t_0)^{3n} = b^3$, hence $(1 + t_0 + \cdots + t_0^{p-1})^{3n} = b^3 t_0^{(p-1)n}$, so t_0 is a zero of a *monic* polynomial of degree $3n(p-1)$ with *integer* coefficients. This is a contradiction, since t_0 is not an algebraic integer.

Avoiding non-trivial solutions to a system of balanced linear equations

As a final application, we briefly look into the problem which will be studied in more detail in the next chapter. Let $\mathbb{F}_q$ be a finite field, and let $A = (a_{ij}) \in \mathbb{F}_q^{m \times k}$ be a fixed matrix over $\mathbb{F}_q$. Consider the linear system

$$\begin{cases} a_{11}\boldsymbol{x_1} + \cdots + a_{1k}\boldsymbol{x_k} = 0, \\ \qquad\qquad \vdots \\ a_{m1}\boldsymbol{x_1} + \cdots + a_{mk}\boldsymbol{x_k} = 0; \end{cases} \tag{$\star$}$$

with variables $\boldsymbol{x_1}, \dots, \boldsymbol{x_k} \in \mathbb{F}_q^n$. We say that $(\star)$ is *balanced* if each of the row sums $a_{i1} + \cdots + a_{ik}$ of the coefficient matrix is equal to zero. If $(\star)$ is balanced, then a solution $(\boldsymbol{x_1}, \dots, \boldsymbol{x_k})$ is called *trivial* if $\boldsymbol{x_1} = \boldsymbol{x_2} = \cdots = \boldsymbol{x_k}$, and *non-trivial* otherwise. When the number of variables is sufficiently large, a straightforward application of the slice rank method shows that a set $S \subseteq \mathbb{F}_q^n$ without non-trivial solutions must have exponentially small density:

Theorem 5.12 ([TS16], [Sau22, Theorem 1.1])**.** *Let q be a prime power, and let $(\star)$ be a balanced linear system over $\mathbb{F}_q$ with k variables and m equations, where $k \geq 2m+1$. Write $C_{q,m,k} := J(q, \frac{(q-1)m}{k}) < q$. Then for all $n \in \mathbb{N}$, every subset $S \subseteq \mathbb{F}_q^n$ without non-trivial solutions to $(\star)$ has size $|S| \leq (C_{q,m,k})^n$.*

Proof. Let $S \subseteq \mathbb{F}_q^n$ be a set without non-trivial solutions to $(\star)$. Define $T : S^k \to \mathbb{F}_q$ by

$$\begin{aligned} T(\boldsymbol{x_1}, \dots, \boldsymbol{x_k}) &:= \prod_{i=1}^{m} \prod_{j=1}^{n} (1 - (a_{i1}x_{1j} + \cdots + a_{ik}x_{kj})^{q-1}) \\ &= \begin{cases} 1, & \text{if } (\boldsymbol{x_1}, \dots, \boldsymbol{x_k}) \text{ is a solution to } (\star); \\ 0, & \text{otherwise;} \end{cases} \\ &= \begin{cases} 1, & \text{if } \boldsymbol{x_1} = \boldsymbol{x_2} = \cdots = \boldsymbol{x_k}; \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

By Proposition 5.3, we have $\mathrm{sr}(T) = |S|$. On the other hand, by Lemma 5.4 and Lemma 5.5, we have $\mathrm{sr}(T) \leq k \cdot m_{q,n,(q-1)\frac{nm}{k}} \leq k \cdot J(q, (q-1)\frac{m}{k})^n = k \cdot (C_{q,m,k})^n$, where the second inequality relies on the assumption that $\frac{m}{k} < \frac{1}{2}$. This shows that $|S| \leq k \cdot (C_{q,m,k})^n$.

To get rid of the additional factor k, we use the power trick. For every $\ell \in \mathbb{N}$, the set $S^\ell \subseteq (\mathbb{F}_q^n)^\ell \cong \mathbb{F}_q^{n\ell}$ also has no non-trivial solutions to $(\star)$. Therefore we have $|S^\ell| \leq k \cdot (C_{q,m,k})^{n\ell}$, hence $|S| \leq k^{1/\ell} \cdot (C_{q,m,k})^n$. Letting $\ell \to \infty$, we conclude that $|S| \leq (C_{q,m,k})^n$. ∎

Since a 3-term arithmetic progression can be encoded by a single balanced linear equation in three variables, Theorem 5.12 contains Theorem 5.8 as a special case.

It is important to note that Theorem 5.12 only applies to systems with sufficiently many variables. The case where $k \leq 2m$ appears to be out of reach for current (slice rank) methods. In the next chapter, we will extend Theorem 5.12 in another direction, by looking at the problem of finding/avoiding solutions of higher non-degeneracy.

Chapter 6

Avoiding solutions to a system of balanced linear equations

The solution of the cap set problem shows that a subset $S \subseteq \mathbb{F}_p^n$ without 3-term arithmetic progressions must have exponentially small density. For k-term arithmetic progressions ($k \geq 4$), this problem is wide open, and is believed to be beyond the reach of current slice rank methods. In this chapter, we study an application of the slice rank method to the broader problem of avoiding non-degenerate solutions to a system of balanced linear equations over $\mathbb{F}_q$.

This chapter is based on the paper [DG21], and is joint work with Dion Gijswijt.

6.1 Introduction

The cap set problem occurs as a special case of several other open problems. Therefore we should ask if the slice rank method can also be used to solve these more general problems. One such problem is to determine whether or not for all values of $3 \leq k \leq p$ there is a constant $c_{p,k} < p$ such that every set $S \subseteq \mathbb{F}_p^n$ with $|S| \geq c_{p,k}^n$ contains a k-term arithmetic progression. For $k = 3$ this is settled by the slice rank method (see [EG17]), but for $k \geq 4$ the problem is wide open. This is believed to be beyond the reach of current slice rank methods.

Instead, research has shifted to other related problems. Recently, Mimura and Tokushige [MT19a, MT19b, MT20] and Sauermann [Sau22] started developing techniques to bound the maximum size of a subset of $\mathbb{F}_q^n$ which avoids non-degenerate solutions to a given system of linear equations over a finite field $\mathbb{F}_q$. Since a k-term arithmetic progression can be encoded as a system of $k-2$ linear equations, this contains the problem of avoiding k-APs as a special case.

Given a fixed matrix $A = (a_{ij}) \in \mathbb{F}_q^{m \times k}$, we want to bound the maximum size of a subset $S \subseteq \mathbb{F}_q^n$ for which there are no k-tuples $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}) \in S^k$ satisfying

$$\begin{cases} a_{11}\boldsymbol{x_1} + \cdots + a_{1k}\boldsymbol{x_k} = 0, \\ \qquad\qquad \vdots \\ a_{m1}\boldsymbol{x_1} + \cdots + a_{mk}\boldsymbol{x_k} = 0; \end{cases} \tag{$\star$}$$

except possibly trivial/degenerate solutions (more on that later). Note that the variables $\boldsymbol{x_1}, \dots, \boldsymbol{x_k}$ are not taken from $\mathbb{F}_q$, but from $\mathbb{F}_q^n$ as $n \to \infty$.

If $a_{i1} + \dots + a_{ik} \neq 0$ for some i (i.e. the coefficients in one of the rows do not sum to zero), then there are large subsets of $\mathbb{F}_q^n$ with no solutions at all to $(\star)$. Indeed, let $S \subseteq \mathbb{F}_q^n$ be the set of all vectors whose first coordinate is equal to 1. If some row of $(\star)$ does not sum to zero, then S does not contain solutions to $(\star)$, and $|S| = q^{n-1} = \frac{1}{q} \cdot |\mathbb{F}_q^n|$, so S contains a constant proportion of the vectors in $\mathbb{F}_q^n$. (This example is due to Sauermann [Sau22].)

We will henceforth assume that $a_{i1} + \dots + a_{ik} = 0$ for all i. Such equations are called *balanced linear equations* (or *affine dependences*), and the system $(\star)$ is also called *balanced*. Recent results show that the problem becomes much more interesting in this case.

If the system $(\star)$ is balanced, then every set $S \subseteq \mathbb{F}_q^n$ has at least $|S|$ solutions to $(\star)$, namely the solutions of the form $(a, \dots, a)$ for $a \in S$. So the question is: how large does S have to be to guarantee the existence of solutions to $(\star)$ which are somehow non-degenerate? For this we consider three different notions of non-degeneracy:

Definition 6.1. A solution $(\boldsymbol{x_1}, \dots, \boldsymbol{x_k}) \in (\mathbb{F}_q^n)^k$ of $(\star)$ is called:

(a) *non-trivial* if $\boldsymbol{x_1}, \dots, \boldsymbol{x_k}$ are not all equal.

(b) a $(\star)$-*shape*[1] if $\boldsymbol{x_1}, \dots, \boldsymbol{x_k}$ are pairwise distinct.

(c) *generic*[2] if every balanced linear equation (over $\mathbb{F}_q$) satisfied by $(\boldsymbol{x_1}, \dots, \boldsymbol{x_k})$ is a linear combination of the equations in $(\star)$.

The requirements get stronger in each step, moving from (a) to (c). Indeed, it is clear that every $(\star)$-shape is a non-trivial solution. Furthermore, if the system $(\star)$ does not rule out the existence of $(\star)$-shapes in $\mathbb{F}_q^n$ (in other words, if no linear combination of the rows of $(\star)$ forces $\boldsymbol{x_i} = \boldsymbol{x_j}$ for $i \neq j$), then every generic solution is a $(\star)$-shape.

The easiest of these problems is finding a non-trivial solution. If the number of variables is sufficiently large (specifically, if $k \geq 2m + 1$), then this can be done by a routine application of the slice rank method, as we showed in the previous chapter:

Theorem 6.2 (see Theorem 5.12)**.** *If $k \geq 2m + 1$, then there exists a constant $\Gamma_{q,m,k} < q$ such that every subset $S \subseteq \mathbb{F}_q^n$ of size at least $(\Gamma_{q,m,k})^n$ has a non-trivial solution of $(\star)$.*

If $k \leq 2m$, then the problem is believed to be beyond the reach of current (slice rank) methods. Therefore we will assume throughout this chapter that $k \geq 2m+1$. The aim of this chapter is to refine Theorem 6.2 to the stronger notions of non-degeneracy from Definition 6.1. For this we use the following terminology:

Definition 6.3. The linear system $(\star)$ is called:

(a) *moderate*[1] if there exist constants $\beta, \gamma > 0$ with $\gamma < q$ such that every subset $S \subseteq \mathbb{F}_q^n$ of size at least $\beta \cdot \gamma^n$ contains a $(\star)$-shape;

[1] Following terminology from Mimura and Tokushige [MT19a, MT19b, MT20].

[2] Terminology introduced by the authors.

(b) *temperate*[3] if there exist constants $\beta, \gamma > 0$ with $\gamma < q$ such that every subset $S \subseteq \mathbb{F}_q^n$ of size at least $\beta \cdot \gamma^n$ contains a generic solution of $(\star)$.

If $(\star)$ consists of the single equation $\boldsymbol{x_1} + \cdots + \boldsymbol{x_p} = 0$ over $\mathbb{F}_p$ (with p prime), then the existence of $(\star)$-shapes is tightly linked to the Erdős–Ginzburg–Ziv constant of the group $\mathbb{F}_p^n$. If $p \geq 3$, then this system is moderate over $\mathbb{F}_p$; this is implicit in [Nas20a] and [Sau21]. Furthermore, the method in [Sau21] can be easily adapted to show that every balanced linear equation with at least 3 variables forms a moderate linear system.

The problem of determining whether or not a system of two or more equations is moderate was first studied by Mimura and Tokushige [MT19a, MT19b, MT20].[4] They showed that several specific linear systems are moderate. Although all of their proofs rely on more or less the same idea, the details of the proofs are so different that a new proof was needed for each new system. We discuss some of their results in more detail in §6.7.

The first general result in this direction was found by Sauermann [Sau22]. In an elaborate proof, using a new application of the slice rank method and a subspace sampling argument, she showed that $(\star)$-shapes can always be found if the number of variables is sufficiently large and if the system is very much non-degenerate:

Theorem 6.4 ([Sau22, Theorem 1.2])**.** *If $k \geq 3m$ and every $m \times m$ submatrix of A is invertible, then $(\star)$ is moderate.*

Despite its generality, this result does not replace the results of Mimura and Tokushige, because the systems they studied have many singular $m \times m$ submatrices (so Theorem 6.4 does not apply).

The third and final problem is that of finding a generic solution. A partial result in this direction was found by Sauermann, who showed that solutions of higher dimension exist as the number of variables becomes larger:

Theorem 6.5 ([Sau22, Theorem 1.3])**.** *If $r \geq 2$ and $k \geq 2m - 1 + r$, then there are constants $C^{\mathrm{rank}}_{p,m,k,r} \geq 1$ and $\Gamma^{\mathrm{rank}}_{p,m,k,r} < p$ such that every subset $S \subseteq \mathbb{F}_p^n$ of size at least $C^{\mathrm{rank}}_{p,m,k,r} \cdot (\Gamma^{\mathrm{rank}}_{p,m,k,r})^n$ has a solution $(\boldsymbol{x_1}, \dots, \boldsymbol{x_k}) \in S^k$ of $(\star)$ satisfying $\dim(\mathrm{span}(\boldsymbol{x_1}, \dots, \boldsymbol{x_k})) \geq r$.*

Finding solutions of high dimension is closely related to finding a generic solution, as we explain in §6.5.

Main results of this chapter

The main results of this chapter are twofold. First, we prove a general result on finding $(\star)$-shapes, which contains most of the results from [MT19a, MT19b, MT20] as special cases. Second, we prove a general result for finding generic solutions, which we believe to be the first of its kind. We should point out that these results have since been superseded by an even more general result of Gijswijt [Gij21].

[3] Terminology introduced by the authors.

[4] Similar results over the integers had been obtained by Ruzsa in the 1990s [Ruz93, Ruz95], but Mimura and Tokushige were the first to study this problem for vector spaces over a finite field.

Throughout the chapter, we focus on a specific class of systems that is completely different from the class of systems studied by Sauermann. Where Sauermann's result (Theorem 6.4 above) requires every $m \times m$ submatrix to be invertible, we require the opposite: there must be sufficiently many linear dependencies between the columns. Specifically, we focus on the class of 'type (RC)' linear systems, which we define as follows:

Definition 6.6. Consider the linear system $(\star)$, whose coefficients are specified by the matrix $A = (a_{ij}) \in \mathbb{F}_q^{m \times k}$.

(a) We say that two indices in $[k]$ are *equivalent* if the corresponding columns of A are nonzero scalar multiples of one another. This defines an equivalence relation on $[k]$. We will refer to the equivalence classes of this equivalence relation as the *column equivalence classes*.

(b) We say that $(\star)$ is *a type (RC) linear system*[5] if it is balanced and has at most one column equivalence class of size 1.

(c) We say that a column equivalence class *sums to zero* if the columns indexed by that class add up to the zero vector.

Examples of type (RC) linear systems will be given in §6.7 below. Among these examples are the systems studied by Mimura and Tokushige.

The assumptions made throughout this chapter can be summarized as follows:

Situation 6.7. Let $(\star)$ be a type (RC) linear system, given by the coefficient matrix $A = (a_{ij}) \in \mathbb{F}_q^{m \times k}$, with ℓ column equivalence classes. Furthermore, assume that $(\star)$ is non-degenerate and irreducible (see Definition 6.13 below).

In all of our main results below, we assume that $(\star)$ and A are as in Situation 6.7. In particular, we always assume that $(\star)$ is irreducible. However, we note that our results can also be applied to reducible systems. We show in Proposition 6.14 (resp. Proposition 6.27) that a system is moderate (resp. temperate) if and only if every irreducible subsystem is moderate (resp. temperate).

Our first main result is a sufficient condition for a type (RC) linear sytem to be moderate.

Theorem 6.8. *Let $(\star)$, A, m, k and ℓ be as in Situation 6.7. Suppose that $(\star)$ satisfies at least one of the following additional properties:*

(i) *none of the column equivalence classes of size 2 sums to zero;*

(ii) *every column equivalence class sums to zero, and $k \geq 3$.*

Then $(\star)$ is moderate.

[5]Terminology introduced by the authors ('RC' stands for 'repeated columns').

This result encompasses most of the systems studied by Mimura and Tokushige, and the rest can be recovered using a slight modification of our proof. See §6.7 for a detailed discussion.

Our second main result is a sufficient condition for a type (RC) linear sytem to be temperate.

Theorem 6.9. *Let $(\star)$, A, m, k and ℓ be as in Situation 6.7. Suppose that $(\star)$ satisfies at least one of the following additional properties:*

(i) *none of the column equivalence classes sums to zero, and $\ell = m + 1$;*

(ii) *every column equivalence class sums to zero.*

Then $(\star)$ is temperate.

The requirements of Theorem 6.9 are more restrictive than those of Theorem 6.8.[6] In particular, one of the systems studied by Mimura and Tokushige does not meet these requirements (see §6.7 for a detailed discussion).

We do not know if every irreducible linear system of type (RC) is moderate and/or temperate, but we have the following partial result. We say that a balanced linear equation satisfied by $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}) \in S^k$ *preserves the column equivalence classes of* $(\star)$ if appending that equation to the system $(\star)$ preserves the column equivalence classes. We prove the following:

Theorem 6.10. *Let $(\star)$, A, m, k and ℓ be as in Situation 6.7. Then there exist constants $\beta, \gamma > 0$ with $\gamma < q$ such that every subset $S \subseteq \mathbb{F}_q^n$ of size at least $\beta \cdot \gamma^n$ has a solution $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}) \in S^k$ of $(\star)$ with the following properties:*

(i) *every balanced linear equation satisfied by $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k})$ preserves the column equivalence classes of $(\star)$;*

(ii) $\dim(\mathrm{aff}(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k})) \geq \min(k - \ell, k - 2)$.

Theorem 6.10 improves upon Theorem 6.5 whenever $2 \leq \ell < 2m$; see Remark 6.35.

Finally, we turn to an application of our techniques and results. In characteristic 0, results like Bourgain's theorem [Bou90] (see also [TV06, Chapter 12]) show that it is substantially easier to find long arithmetic progressions in sum sets than in general sets. Using the techniques from this chapter, we establish a similar result in vector spaces over $\mathbb{F}_q$.

Given sets $S_1, \ldots, S_l \subseteq \mathbb{F}_q^n$, we define the *affinely independent restricted sum set* (or *AIR-sumset*) as follows:

$$S_1 \underset{\text{aff}}{\dot{+}} \cdots \underset{\text{aff}}{\dot{+}} S_l := \{\boldsymbol{x_1} + \cdots + \boldsymbol{x_l} \mid \boldsymbol{x_1} \in S_1, \ldots, \boldsymbol{x_l} \in S_l \text{ affinely independent}\}.$$

Further, if $(\star)$ is linear system which is not necessarily balanced, then we say that a solution $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}) \in (\mathbb{F}_q^n)^k$ is *linearly generic* if every linear equation (over $\mathbb{F}_q$)

[6] Except that Theorem 6.9(ii) does not have the condition $k \geq 3$. That condition is included in Theorem 6.8 to rule out the system $\boldsymbol{x_1} - \boldsymbol{x_2} = 0$. It is not hard to see that this particular system is temperate but not moderate.

satisfied by $(\boldsymbol{x_1}, \dots, \boldsymbol{x_k})$ is a linear combination of the equations in $(\star)$. By comparison, the solutions which we call *generic* throughout this chapter (see Definition 6.1(c)) only satisfy this property for *balanced* linear equations (so by 'generic' we will always mean 'affinely generic').

Corollary 6.11. *Let $\mathbb{F}_q$ be a finite field, let $(\star)$ be a (not necessarily balanced) linear system over $\mathbb{F}_q$, and let $c_1, \dots, c_l \in \mathbb{F}_q \setminus \{0\}$ with $c_1 + \dots + c_l = 0$. Then there are constants $\beta, \gamma \geq 1$ with $\gamma < q$ such that, for every subset $S \subseteq \mathbb{F}_q^n$ of size at least $\beta \cdot \gamma^n$, the set $(c_1 \cdot S \underset{\text{aff}}{+} \dots \underset{\text{aff}}{+} c_l \cdot S) \cup \{0\}$ contains a linearly generic solution of $(\star)$.*

Note that Corollary 6.11 does not impose any restriction on the linear system $(\star)$; that is, the coefficient matrix $A \in \mathbb{F}_q^{m \times k}$ can be arbitrary. This is a significant difference with our main results and Sauermann's result (Theorem 6.4 above), which only work for very specific classes of linear systems.

In Corollary 6.11, we only need to append 0 to the AIR-sumset when one of the single-variable equations $\boldsymbol{x_j} = 0$ $(j \in [k])$ can be written as a linear combination of the equations in the linear system $(\star)$. If this is not the case, then a linearly generic solution $(\boldsymbol{x_1}, \dots, \boldsymbol{x_k})$ will satisfy $\boldsymbol{x_j} \neq 0$ for all $j \in [k]$, so it is not necessary to append 0 to the AIR-sumset.

By letting $(\star)$ be the system that encodes a k-term arithmetic progression, Corollary 6.11 contains the following special case:

Corollary 6.12. *Let p be prime, and let $3 \leq k \leq p$. Then, for every subset $S \subseteq \mathbb{F}_p^n$ of size at least $p^{1+(1-\frac{1}{k})n}$, the set $(S - S) \setminus \{0\}$ contains a non-trivial k-term arithmetic progression.*

We note that this special case can be proved without using the slice rank method, using only a simple counting argument (see §6.7 for details).

Overview of the main ideas and organization of this chapter

Main ideas. There are two new techniques in this chapter.

First, the majority of our results depend on a 'replacement trick'. This trick works roughly as follows. If the j_1-th and j_2-th columns of A are non-zero multiples of one another, and if we have a long enough list $\{(\boldsymbol{x_1^{(i)}}, \dots, \boldsymbol{x_k^{(i)}})\}_{i=1}^{L}$ of pairwise disjoint solutions to $(\star)$, then we use tricoloured sum-free sets to recombine these solutions to obtain new solutions of $(\star)$. This is done by taking one of the solutions from this list, say $(\boldsymbol{x_1^{(i)}}, \dots, \boldsymbol{x_k^{(i)}})$, and replacing $\boldsymbol{x_{j_1}^{(i)}}$ and $\boldsymbol{x_{j_2}^{(i)}}$ by (respectively) $\boldsymbol{x_{j_1}^{(i')}}$ and $\boldsymbol{x_{j_2}^{(i'')}}$, for some $i', i'' \neq i$. We show in Corollary 6.21 that there exists $i \in [L]$ which admits one such replacement (the 'single replacement trick'), and in Corollary 6.30 that there exists $i \in [L]$ which admits many replacements (the 'multiple replacement trick').

The second main ingredient in our proofs is Lemma 6.22, which shows that, for every subset $S \subseteq \mathbb{F}_q^n$ of size at least $q^{1+(1-\frac{1}{k})n}$, the difference set $S - S$ contains linearly generic solutions to every linear system in k variables. The proof relies only on a simple counting argument, using the pigeonhole principle.

We point out that this chapter does not make use of the full strength of Theorem 6.2, as we only use the slice rank method for 3-tensors. Indeed, the replacement trick relies

on tricoloured sum-free sets, and Lemma 6.22 does not rely on the slice rank method at all.

The constants. Theorem 6.8(i), Theorem 6.9(i), and Theorem 6.10 rely only on the replacement trick. Hence, the base of the exponent in the upper bounds from these theorems[7] is equal to Γ_q, the constant from the bound on tricoloured sum-free sets (see Theorem 6.17).

Theorem 6.8(ii), Theorem 6.9(ii), and Corollary 6.11 rely on a combination of the replacement trick and Lemma 6.22. Hence, the base of the exponent in the upper bounds from these theorems is the maximum of Γ_q and $q^{\frac{k-1}{k}}$.

Corollary 6.12 relies solely on Lemma 6.22. The base of the exponent in the upper bound is $p^{1-\frac{1}{k}}$.

Organization of the chapter. This chapter consists of three parts.

First, in §6.2–6.4, we focus on moderate systems. In §6.2, we discuss the generalities of moderate systems, and we show that we may restrict our attention to irreducible systems. In §6.3, we establish the 'single replacement trick', and use it to prove Theorem 6.8(i). In §6.4, we establish the other main technique of this chapter (Lemma 6.22), and combine it with the replacement trick to prove Theorem 6.8(ii).

Second, in §6.5–6.6, we focus on temperate systems. In §6.5, we discuss the generalities of temperate systems. Here we show how the problem of finding solutions of high rank is related to the problem of finding a generic solution, and we show that we may once again restrict our attention to irreducible systems. In §6.6, we establish the 'multiple replacement trick', and use it to prove Theorem 6.9 and Theorem 6.10.

Finally, in §6.7, we discuss several examples and applications. Here we prove Corollary 6.11 and Corollary 6.12, and we recover most of the results from [MT19a, MT19b, MT20] as special cases of our results. Furthermore, we show that the system conjectured to be moderate in [MT20] is indeed moderate.

6.2 Preliminaries on moderate systems

In this chapter, we study linear systems of the form

$$\begin{cases} a_{11}\boldsymbol{x_1} + \cdots + a_{1k}\boldsymbol{x_k} = 0, \\ \qquad\qquad \vdots \\ a_{m1}\boldsymbol{x_1} + \cdots + a_{mk}\boldsymbol{x_k} = 0; \end{cases} \tag{$\star$}$$

with coefficient matrix $A = (a_{ij}) \in \mathbb{F}_q^{m\times k}$ and variables $\boldsymbol{x_1}, \ldots, \boldsymbol{x_k} \in \mathbb{F}_q^n$.

Following standard usage, we say that two linear systems $(\star)$ and $(\star')$ are *equivalent* if each equation in $(\star)$ is a linear combination of the equations in $(\star')$ and vice versa. Furthermore, we say that a variable $\boldsymbol{x_i}$ is *used* by the linear system $(\star)$ if it occurs with non-zero coefficient in at least one equation.

[7] By 'the base of the exponent in the upper bound', we mean the constant $\gamma < q$ in the upper bound $\beta \cdot \gamma^n$.

Definition 6.13. The linear system $(\star)$ is said to be:

(a) *non-degenerate* if the rows of A are linearly independent and every variable is used;

(b) *reducible* if it is equivalent to a linear system $(\star')$ with the property that the variables $\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}$ can be partitioned into two or more classes in such a way that every equation in $(\star')$ only uses variables from one partition class. If this is not the case, then $(\star)$ is said to be *irreducible*.

Passing to an equivalent system or deleting columns with only zeroes does not change the problem of finding a $(\star)$-shape, so we may assume without loss of generality that $(\star)$ is non-degenerate. The following proposition shows that we can also restrict our attention to irreducible systems.

Proposition 6.14. *Suppose that $(\star)$ is equivalent to a linear system $(\star')$ whose coefficient matrix can be written as*

$$\begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$$

for some $A_1 \in \mathbb{F}_q^{m_1 \times k_1}$ and $A_2 \in \mathbb{F}_q^{m_2 \times k_2}$ with $m_1, m_2, k_1, k_2 \neq 0$. Then $(\star)$ is moderate if and only if the systems given by A_1 and A_2 are moderate.

Proof. If $(\star')$ is moderate, then it is easy to see that the same holds for the systems given by A_1 and A_2.

Suppose that for $i = 1, 2$, the system given by A_i is moderate, with constants $\beta_i, \gamma_i > 0$, where $\gamma_i < q$. Let $S \subseteq \mathbb{F}_q^n$ be a set of size at least $\max(\beta_1\gamma_1^n, k_1 + \beta_2\gamma_2^n)$. Since $|S| \geq \beta_1\gamma_1^n$, we may choose an A_1-shape $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_{k_1}})$ in S. Then, since $|S \setminus \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_{k_1}}\}| \geq \beta_2\gamma_2^n$, we may choose an A_2-shape in $S \setminus \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_{k_1}}\}$. Since $\max(\beta_1\gamma_1^n, k_1 + \beta_2\gamma_2^n) \in \mathcal{O}(\max(\gamma_1, \gamma_2)^n)$, this shows that $(\star')$, and therefore $(\star)$, is moderate. ∎

Therefore we may restrict our attention to irreducible systems, as stipulated in Situation 6.7.

The following proposition will be useful later on.

Proposition 6.15. *Let $(\star)$ be a linear system given by the matrix $A = (a_{ij}) \in \mathbb{F}_q^{m \times k}$. If $(\star)$ is non-degenerate and irreducible, and if $m \geq 2$, then every non-zero linear equation implied by $(\star)$ uses at least two column equivalence classes, and $\ell \geq m + 1$.*

Proof. Let ℓ be the number of column equivalence classes, and note that $m = \operatorname{rank}(A) \leq \ell$ (recall that the columns with indices in the same column equivalence class are scalar multiples of each other). Suppose for the sake of contradiction that some linear combination of the rows of $(\star)$ uses exactly one column equivalence class. By passing to an equivalent system and permuting the columns, we may assume without loss of generality that the first row of $(\star)$ only uses the column equivalence class $C = \{1, \ldots, |C|\} \subseteq [k]$. Since the columns indexed by C are non-zero multiples of one another, we have $a_{1j} \neq 0$ for all $j \in C$.

By Gaussian elimination, we may pass to an equivalent system $(\star')$, given by the matrix $A' = (a'_{ij}) \in \mathbb{F}_q^{m\times k}$, such that $a_{i1} = 0$ for all $i > 1$. Since elementary row operations preserve the column equivalence classes, we have $a_{ij} = 0$ for all $(i,j) \in \{2,\dots,m\} \times C$. It follows that every row in $(\star')$ uses variables from either C or $[k] \setminus C$, but not both. Since $\ell \geq m \geq 2$, we have $|C|, |[k] \setminus C| \neq 0$, so it follows that $(\star)$ is reducible. This is a contradiction, so we conclude that every (non-zero) equation implied by $(\star)$ uses at least two column equivalence classes.

To prove that $\ell \geq m+1$, let A' be the matrix obtained by deleting from A the columns in one column equivalence class. By the above, every non-zero linear combination of the rows of A' uses at least one of the remaining $\ell-1$ column equivalence classes. It follows that $\operatorname{rank}(A') = m$, so $\ell - 1 \geq m$. ■

6.3 Proof of Theorem 6.8(i)

In this section, we develop the first main technique (the 'single replacement trick', see Corollary 6.21) and use it to prove Theorem 6.8(i).

We recall the definition of tricoloured sum-free sets (already used in §5.3):

Definition 6.16. Let G be an abelian group. A sequence $\{(x_i, y_i, z_i)\}_{i=1}^L$ in G^3 is called a *tricoloured sum-free set in* G if for all $i, i', i'' \in [L]$ one has $x_i + y_{i'} + z_{i''} = 0$ if and only if $i = i' = i''$.

Note that the definition implies $|\{x_1,\dots,x_L\}| = |\{y_1,\dots,y_L\}| = |\{z_1,\dots,z_L\}| = L$; that is, in a tricoloured sum-free set there can be no repetitions in each of the coordinates (separately).

In Corollary 5.10 and Remark 5.11, we proved the following exponential upper bound on the size of tricoloured sum-free sets in $\mathbb{F}_q^n$:

Theorem 6.17 (cf. Corollary 5.10, Remark 5.11)**.** *Let $q = p^s$ be a prime power, and define $\Gamma_q := J(p, \frac{p-1}{3})^s$. Then for every tricoloured sum-free set $\{(\boldsymbol{x_i}, \boldsymbol{y_i}, \boldsymbol{z_i})\}_{i=1}^L$ in $\mathbb{F}_q^n$ one has $L < (\Gamma_q)^n$.*

To prove the 'single replacement trick', we start with the following lemma.

Lemma 6.18. *Let q be a prime power, and let Γ_q be as in Theorem 6.17. Let $\alpha, \beta \in \mathbb{F}_q \setminus \{0\}$, let $\boldsymbol{x_1},\dots,\boldsymbol{x_L} \in \mathbb{F}_q^n$ be distinct, and let $\boldsymbol{y_1},\dots,\boldsymbol{y_L} \in \mathbb{F}_q^n$ be distinct. If $L \geq (\Gamma_q)^n$, then there exist $i, i', i'' \in [L]$ with $i \neq i'$, i'' and $\alpha\boldsymbol{x_i} + \beta\boldsymbol{y_i} = \alpha\boldsymbol{x_{i'}} + \beta\boldsymbol{y_{i''}}$.*

Proof. For $i \in [L]$, define $\boldsymbol{z_i} = \alpha\boldsymbol{x_i} + \beta\boldsymbol{y_i}$. Each triple in the sequence $\{(\alpha\boldsymbol{x_i}, \beta\boldsymbol{y_i}, -\boldsymbol{z_i})\}_{i=1}^L$ sums to zero, but we have $L \geq (\Gamma_q)^n$, so it follows from Theorem 6.17 that this sequence is not a tricoloured sum-free set. Therefore we may choose $i, i', i'' \in [L]$, not all equal, such that $\alpha\boldsymbol{x_i} + \beta\boldsymbol{y_i} = \boldsymbol{z_i} = \alpha\boldsymbol{x_{i'}} + \beta\boldsymbol{y_{i''}}$.

Suppose that $i'' = i$. Then we have $\alpha\boldsymbol{x_i} = \alpha\boldsymbol{x_{i'}}$, hence $\boldsymbol{x_i} = \boldsymbol{x_{i'}}$ (because $\alpha \neq 0$), and therefore $i = i'$ (because $\boldsymbol{x_1},\dots,\boldsymbol{x_L}$ are distinct), contrary to our assumption that i, i' and i'' are not all equal. This is a contradiction, so we must have $i'' \neq i$. An analogous argument shows that $i' \neq i$. ■

Remark 6.19. In Lemma 6.18, we do not require that $i' \neq i''$. The case that $i' = i''$ corresponds to the case that $\boldsymbol{z_1}, \ldots, \boldsymbol{z_L}$ are not all distinct. This does not matter for the rest of the proof.

Definition 6.20. We say that two solutions $\vec{\boldsymbol{x}} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_k})$ and $\vec{\boldsymbol{y}} = (\boldsymbol{y_1}, \ldots, \boldsymbol{y_k})$ to $(\star)$ are *disjoint* if $\{\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}\} \cap \{\boldsymbol{y_1}, \ldots, \boldsymbol{y_k}\} = \varnothing$. Note that we do not require the $\boldsymbol{x_j}$ (resp. the $\boldsymbol{y_j}$) to be pairwise distinct.

Corollary 6.21 ('Single replacement trick')**.** *Let $\{(\boldsymbol{x_1^{(i)}}, \ldots, \boldsymbol{x_k^{(i)}})\}_{i=1}^{L}$ be a list of pairwise disjoint solutions of $(\star)$, and suppose that j_1 and j_2 are distinct indices in the same column equivalence class. If $L \geq (\Gamma_q)^n$, then there exist $i, i', i'' \in [L]$ with $i \neq i', i''$ such that the k-tuple $(\boldsymbol{y_1}, \ldots, \boldsymbol{y_k}) \in (\mathbb{F}_q^n)^k$ given by*

$$\boldsymbol{y_j} = \begin{cases} \boldsymbol{x_j^{(i)}}, & \text{if } j \neq j_1, j_2; \\ \boldsymbol{x_j^{(i')}}, & \text{if } j = j_1; \\ \boldsymbol{x_j^{(i'')}}, & \text{if } j = j_2; \end{cases}$$

is also a solution of $(\star)$.

Proof. Since the j_1-th and j_2-th column of $(\star)$ are multiples of one another, we may choose a vector $\boldsymbol{v} \in \mathbb{F}_q^m$ and constants $\alpha, \beta \neq 0$ such that the j_1-th column is equal to $\alpha\boldsymbol{v}$ and the j_2-th column is equal to $\beta\boldsymbol{v}$.

By assumption, the vectors $\boldsymbol{x_{j_1}^{(1)}}, \ldots, \boldsymbol{x_{j_1}^{(L)}}$ are distinct, and likewise the vectors $\boldsymbol{x_{j_2}^{(1)}}, \ldots, \boldsymbol{x_{j_2}^{(L)}}$ are distinct, so it follows from Lemma 6.18 that there exist $i, i', i'' \in [L]$ with $i \neq i', i''$ and $\alpha\boldsymbol{x_{j_1}^{(i)}} + \beta\boldsymbol{x_{j_2}^{(i)}} = \alpha\boldsymbol{x_{j_1}^{(i')}} + \beta\boldsymbol{x_{j_2}^{(i'')}}$. Hence, the total contribution of $\boldsymbol{x_{j_1}^{(i)}}$ and $\boldsymbol{x_{j_2}^{(i)}}$ to the equations of $(\star)$ is the same as the contribution of $\boldsymbol{x_{j_1}^{(i')}}$ and $\boldsymbol{x_{j_2}^{(i'')}}$. Since $(\boldsymbol{x_1^{(i)}}, \ldots, \boldsymbol{x_k^{(i)}})$ is a solution of $(\star)$, so is $(\boldsymbol{y_1}, \ldots, \boldsymbol{y_k})$. ∎

We now prove the first main result of this chapter, using the replacement trick from the preceding corollary.

Proof of Theorem 6.8(i). Let $(\star)$, A, m, k and ℓ be as in Situation 6.7, and suppose that $(\star)$ satisfies property (i) from Theorem 6.8 (none of the column equivalence classes of size 2 sums to zero). Furthermore, let Γ_q be the constant from Theorem 6.17.

We prove by induction on λ that, for every $\lambda \in [k]$, there is a constant $\beta_\lambda \geq 1$ such that every subset $S \subseteq \mathbb{F}_q^n$ of size at least $\beta_\lambda \cdot (\Gamma_q)^n$ contains a solution $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}) \in S^k$ of $(\star)$ with at least λ different vectors; that is, $|\{\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}\}| \geq \lambda$. For $\lambda = 1$, this is trivially true with $\beta_1 = 1$, since $(\boldsymbol{x}, \ldots, \boldsymbol{x})$ is a solution of $(\star)$ for every $\boldsymbol{x} \in \mathbb{F}_q^n$.

For the induction step, suppose that $\lambda_0 \in [k-1]$ is given such that the statement is true for $\lambda = \lambda_0$. Define $\beta_{\lambda_0+1} := \beta_{\lambda_0} + P(k, \lambda_0) \cdot k$, where $P(k, \lambda_0)$ denotes the number of partitions of a k-element set into λ_0 parts.

Let $S \subseteq \mathbb{F}_q^n$ be a set of size at least $\beta_{\lambda_0+1} \cdot (\Gamma_q)^n = \beta_{\lambda_0} \cdot (\Gamma_q)^n + P(k, \lambda_0) \cdot (\Gamma_q)^n \cdot k$. Create a list of disjoint solutions $\{(\boldsymbol{x_1^{(i)}}, \ldots, \boldsymbol{x_k^{(i)}})\}_{i=1}^{L_0}$ of $(\star)$ in S, each with at least λ_0 different vectors, by repeatedly finding such a solution in S and removing it from

S. By the induction hypothesis, we can find a new solution as long as the remaining set has size at least $\beta_{\lambda_0} \cdot (\Gamma_q)^n$, and in each step we remove at most k vectors from S, so we find a list of length $L_0 \geq P(k, \lambda_0) \cdot (\Gamma_q)^n$.

If one of the solutions in the list has strictly more than λ_0 different vectors, then we are done. So we may assume that every solution in the list has exactly λ_0 different vectors.

We sort the entries in the list according to their partition pattern. We say that a solution $(\boldsymbol{x_1^{(i)}}, \ldots, \boldsymbol{x_k^{(i)}})$ is *compatible* with a partition $[k] = J_1 \cup \cdots \cup J_{\lambda_0}$ if for all $j_1, j_2 \in [k]$ we have: $\boldsymbol{x_{j_1}^{(i)}} = \boldsymbol{x_{j_2}^{(i)}}$ if and only if j_1 and j_2 belong to the same partition class. Evidently every solution is compatible with exactly one partition. By the pigeonhole principle, we may choose a partition $[k] = J_1 \cup \cdots \cup J_{\lambda_0}$ that occurs at least $(\Gamma_q)^n$ times in our list of solutions. Thus, we obtain a list $\{(\boldsymbol{y_1^{(i)}}, \ldots, \boldsymbol{y_k^{(i)}})\}_{i=1}^{L_1}$ of solutions of the same partition type, where $L_1 \geq (\Gamma_q)^n$.

Now we have two competing partitions of $[k]$, given by the column equivalence classes and the (now fixed) partition type $[k] = J_1 \cup \cdots \cup J_{\lambda_0}$. For $j_1, j_2 \in [k]$, we write $j_1 \parallel j_2$ if j_1 and j_2 are in the same column equivalence class, and $j_1 \equiv j_2$ if j_1 and j_2 belong to the same class in the partition $[k] = J_1 \cup \cdots \cup J_{\lambda_0}$ (i.e. if $\boldsymbol{y_{j_1}^{(i)}} = \boldsymbol{y_{j_2}^{(i)}}$ for all $i \in [L_1]$).

Since $\lambda_0 < k$, we may choose distinct $j_0, j_1 \in [k]$ with $j_0 \equiv j_1$. Furthermore, since $(\star)$ has at most one column equivalence class of size 1, we may assume without loss of generality that j_1 belongs to a column equivalence class of size 2 or more. We distinguish two cases, depending on which of the column equivalence classes j_0 and j_1 belong to.

- **Case 1:** $j_0 \nparallel j_1$ or j_0 and j_1 belong to the same column equivalence class of size at least 3. In this case, we may choose $j_2 \neq j_0, j_1$ such that $j_1 \parallel j_2$. By Corollary 6.21, there is a solution $(\boldsymbol{z_1}, \ldots, \boldsymbol{z_k})$ of $(\star)$ of the form

$$\boldsymbol{z_j} = \begin{cases} \boldsymbol{y_j^{(i)}}, & \text{if } j \neq j_1, j_2; \\ \boldsymbol{y_j^{(i')}}, & \text{if } j = j_1; \\ \boldsymbol{y_j^{(i'')}}, & \text{if } j = j_2; \end{cases}$$

  for some $i, i', i'' \in [L_1]$ with $i \neq i', i''$. In other words, $(\boldsymbol{z_1}, \ldots, \boldsymbol{z_k})$ is obtained by taking the solution $(\boldsymbol{y_1^{(i)}}, \ldots, \boldsymbol{y_k^{(i)}})$ and replacing two entries.

  We prove that $|\{\boldsymbol{z_1}, \ldots, \boldsymbol{z_k}\}| \geq \lambda_0 + 1$. First, note that $\{\boldsymbol{z_{j_1}}, \boldsymbol{z_{j_2}}\} \cap \{\boldsymbol{z_j} \mid j \neq j_1, j_2\} = \varnothing$, since the solutions in the list were disjoint. Now we distinguish two cases.

  - If $j_1 \equiv j_2$, then the removal of the j_1-th and j_2-th vectors from $(\boldsymbol{y_1^{(i)}}, \ldots, \boldsymbol{y_k^{(i)}})$ does not change the number of different vectors, since $\boldsymbol{y_{j_0}^{(i)}} = \boldsymbol{y_{j_1}^{(i)}} = \boldsymbol{y_{j_2}^{(i)}}$. We replace them by two vectors $\boldsymbol{z_{j_1}}, \boldsymbol{z_{j_2}}$ which are distinct from the other vectors in the solution (but possibly $\boldsymbol{z_{j_1}} = \boldsymbol{z_{j_2}}$), so the number of different vectors increases by at least 1.

– If $j_1 \not\equiv j_2$, then the removal of j_1-th and j_2-th vectors from $(\boldsymbol{y}_1^{(i)}, \ldots, \boldsymbol{y}_k^{(i)})$ decreases the number of different vectors by at most 1, because $\boldsymbol{y}_{j_0}^{(i)} = \boldsymbol{y}_{j_1}^{(i)}$. In this case we are guaranteed to have $\boldsymbol{z}_{j_1} \neq \boldsymbol{z}_{j_2}$: different solutions in the list are disjoint, but even within the same solution the j_1-th and j_2-th entry are always different (because $j_1 \not\equiv j_2$). Thus, adding $\boldsymbol{z}_{j_1}$ and $\boldsymbol{z}_{j_2}$ to the solution increases the number of different vectors by 2. The net effect is an increase of at least 1.

This proves our claim that $|\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k\}| \geq \lambda_0 + 1$.

- **Case 2:** j_0 and j_1 belong to the same column equivalence class of size 2. Then, by assumption (i) from the theorem statement, the j_0-th and j_1-th columns of $(\star)$ do not sum to zero.

  By Corollary 6.21, there is a solution $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k)$ of $(\star)$ of the form

$$\boldsymbol{z}_j = \begin{cases} \boldsymbol{y}_j^{(i)}, & \text{if } j \neq j_0, j_1; \\ \boldsymbol{y}_j^{(i')}, & \text{if } j = j_0; \\ \boldsymbol{y}_j^{(i'')}, & \text{if } j = j_1; \end{cases}$$

  for some $i, i', i'' \in [L_1]$ with $i \neq i', i''$.

  Suppose for the sake of contradiction that $\boldsymbol{z}_{j_0} = \boldsymbol{z}_{j_1}$; that is, $\boldsymbol{y}_{j_0}^{(i')} = \boldsymbol{y}_{j_1}^{(i'')}$. Since the j_0-th and j_1-th columns of $(\star)$ do not sum to zero, and since $\boldsymbol{y}_{j_0}^{(i)} = \boldsymbol{y}_{j_1}^{(i)}$, the fact that both $(\boldsymbol{y}_1^{(i)}, \ldots, \boldsymbol{y}_k^{(i)})$ and $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k)$ are solutions of $(\star)$ implies that $\boldsymbol{y}_{j_0}^{(i')} = \boldsymbol{y}_{j_1}^{(i'')} = \boldsymbol{y}_{j_0}^{(i)} = \boldsymbol{y}_{j_1}^{(i)}$. This is a contradiction, because $i \neq i', i''$, and different solutions of the list are disjoint. Therefore we must have $\boldsymbol{z}_{j_0} \neq \boldsymbol{z}_{j_1}$.

  The removal of $\boldsymbol{y}_{j_0}^{(i)}$ and $\boldsymbol{y}_{j_1}^{(i)}$ from the solution decreases the number of different vectors by at most 1, since $\boldsymbol{y}_{j_0}^{(i)} = \boldsymbol{y}_{j_1}^{(i)}$. On the other hand, putting back $\boldsymbol{z}_{j_0}$ and $\boldsymbol{z}_{j_1}$ increases the number of different vectors by 2, since we have $\boldsymbol{z}_{j_0} \neq \boldsymbol{z}_{j_1}$ and $\{\boldsymbol{z}_{j_1}, \boldsymbol{z}_{j_2}\} \cap \{\boldsymbol{z}_j \mid j \neq j_1, j_2\} = \varnothing$. The net effect is an increase of at least 1, so we have $|\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k\}| \geq \lambda_0 + 1$. ■

6.4 Proof of Theorem 6.8(ii)

In this section, we develop our second main technique (Lemma 6.22) and combine it with the techniques from the previous section to prove Theorem 6.8(ii).

Lemma 6.22. *Let $A = (a_{ij}) \in \mathbb{F}_q^{m \times k}$ be a non-zero matrix and let $S \subseteq \mathbb{F}_q^n$ have size at least $q^{1+(1-\frac{1}{k})n}$. Then there are $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k), (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k) \in S^k$ such that, for all $\boldsymbol{b} = (b_1, \ldots, b_k) \in \mathbb{F}_q^k$, one has $b_1\boldsymbol{x}_1 + \cdots + b_k\boldsymbol{x}_k = b_1\boldsymbol{y}_1 + \cdots + b_k\boldsymbol{y}_k$ if and only if $\boldsymbol{b}$ is a linear combination of the rows of A.*

Proof. By removing redundant rows, we may assume without loss of generality that $\operatorname{rank} A = m$. If $k = m$, then we can take $\boldsymbol{x} = \boldsymbol{y} \in S^k$ arbitrary. Hence, we may assume

that $k \geq m+1$. By performing elementary row operations and permuting columns, we may assume without loss of generality that A is of the form $[A'\ I_m]$ for some $A' \in \mathbb{F}_q^{m\times(k-m)}$.

The matrix A defines a function $f : (\mathbb{F}_q^n)^k \to (\mathbb{F}_q^n)^m$, where $[f(\boldsymbol{x_1},\dots,\boldsymbol{x_k})]_i = a_{i1}\boldsymbol{x_1}+\cdots+a_{ik}\boldsymbol{x_k}$. By the pigeonhole principle, we may choose some $\vec{\boldsymbol{z}} = (\boldsymbol{z_1},\dots,\boldsymbol{z_m}) \in (\mathbb{F}_q^n)^m$ such that the set $T := f^{-1}(\vec{\boldsymbol{z}}) \cap S^k$ has size $|T| \geq |S|^k/q^{mn} \geq q^k q^{(k-m-1)n}$.

Let $\pi : (\mathbb{F}_q^n)^k \to (\mathbb{F}_q^n)^{k-m}$ be the projection onto the first $k-m$ coordinates, let $g : T \to (\mathbb{F}_q^n)^{k-m}$ be the restriction of π to T, and let $T' := g[T]$. Since A is of the form $[A'\ I_m]$, it is easy to see that for every $(\boldsymbol{x_1},\dots,\boldsymbol{x_{k-m}}) \in (\mathbb{F}_q^n)^{k-m}$ there is exactly one possible choice of $(\boldsymbol{x_{k-m+1}},\dots,\boldsymbol{x_k}) \in (\mathbb{F}_q^n)^m$ such that $f(\boldsymbol{x_1},\dots,\boldsymbol{x_k}) = \vec{\boldsymbol{z}}$. Therefore g is injective, and it follows that $|T'| = |T|$.

Let $D = \{(\boldsymbol{z_1},\dots,\boldsymbol{z_{k-m}}) \in (\mathbb{F}_q^n)^{k-m} \mid \boldsymbol{z_1},\dots,\boldsymbol{z_{k-m}} \text{ are linearly dependent}\}$. Then $|D| < q^{k-m}q^{(k-m-1)n}$ since there are fewer than q^{k-m} possible linear relations.

Choose some $\vec{\boldsymbol{y}}' = (\boldsymbol{y'_1},\dots,\boldsymbol{y'_{k-m}}) \in T'$. Since $|T' - \vec{\boldsymbol{y}}'| = |T'| > |D|$, we have $(T - \vec{\boldsymbol{y}}') \setminus D \neq \varnothing$, so we may choose $(\boldsymbol{x'_1},\dots,\boldsymbol{x'_{k-m}}) \in T'$ such that $\boldsymbol{x'_1} - \boldsymbol{y'_1},\dots,\boldsymbol{x'_{k-m}} - \boldsymbol{y'_{k-m}}$ are linearly independent. Let $(\boldsymbol{x_1},\dots,\boldsymbol{x_k}),(\boldsymbol{y_1},\dots,\boldsymbol{y_k}) \in T \subseteq S^k$ be the (unique) preimages of $(\boldsymbol{x'_1},\dots,\boldsymbol{x'_{k-m}})$ and $(\boldsymbol{y'_1},\dots,\boldsymbol{y'_{k-m}})$ under g. Note that $(\boldsymbol{x_1},\dots,\boldsymbol{x_{k-m}}) = (\boldsymbol{x'_1},\dots,\boldsymbol{x'_{k-m}})$ and $(\boldsymbol{y_1},\dots,\boldsymbol{y_{k-m}}) = (\boldsymbol{y'_1},\dots,\boldsymbol{y'_{k-m}})$, since g is just a coordinate projection.

We claim that $(\boldsymbol{x_1},\dots,\boldsymbol{x_k})$ and $(\boldsymbol{y_1},\dots,\boldsymbol{y_k})$ satisfy the required property.

Since $f(\boldsymbol{x_1},\dots,\boldsymbol{x_k}) = f(\boldsymbol{y_1},\dots,\boldsymbol{y_k}) = \vec{\boldsymbol{z}}$, it is clear that $b_1\boldsymbol{x_1}+\cdots+b_k\boldsymbol{x_k} = b_1\boldsymbol{y_1}+\cdots+b_k\boldsymbol{y_k}$ whenever $(b_1,\dots,b_k)$ is a linear combination of the rows of A.

Now let $\boldsymbol{b} = (b_1,\dots,b_k) \in \mathbb{F}_q^k$ be an arbitrary row vector such that $b_1\boldsymbol{x_1}+\cdots+b_k\boldsymbol{x_k} = b_1\boldsymbol{y_1}+\cdots+b_k\boldsymbol{y_k}$. Since A is of the form $[A'\ I_m]$, we can add a linear combination of the rows of A to $\boldsymbol{b}$ to obtain a vector $\boldsymbol{c} = (c_1,\dots,c_k) \in \mathbb{F}_q^k$ with $c_{k-m+1} = \cdots = c_k = 0$. By linearity, we have $c_1\boldsymbol{x_1}+\cdots+c_k\boldsymbol{x_k} = c_1\boldsymbol{y_1}+\cdots+c_k\boldsymbol{y_k}$, or equivalently,

$$c_1(\boldsymbol{x_1}-\boldsymbol{y_1})+\cdots+c_{k-m}(\boldsymbol{x_{k-m}}-\boldsymbol{y_{k-m}}) = 0.$$

Since $\boldsymbol{x_1}-\boldsymbol{y_1},\dots,\boldsymbol{x_{k-m}}-\boldsymbol{y_{k-m}}$ are linearly independent, it follows that $c_1 = \cdots = c_{k-m} = 0$, so we have $c_j = 0$ for all $j \in [k]$. This shows that $\boldsymbol{b}$ is a linear combination of the rows of A. ■

We now come to the proof of Theorem 6.8(ii). The proof is largely analogous to the proof of Theorem 6.8(i) (see §6.3), the main difference being that we now use Lemma 6.22 to control column equivalence classes that sum to zero.

We prove the following slightly stronger theorem.

Theorem 6.23. *Let $(\star)$, A, m, k and ℓ be as in Situation 6.7. Suppose that there is a partition $[k] = P_1 \cup \cdots \cup P_{2s}$ such that:*

(i) *for all $r \in [s]$, the columns of A indexed by $P_r \cup P_{s+r}$ sum to zero;*

(ii) *if* $(b_1, \dots, b_k) \in \mathbb{F}_q^k \setminus \{0\}$ *is a non-zero linear combination of the rows of* A*, then one has* $\sum_{j \in P_r} b_j \neq 0$ *for at least two different values of* $r \in [s]$*.*[8]

(iii) *if* C *is a column equivalence class of size* 2 *that sums to zero, then there is some* $r \in [s]$ *such that* $C = P_r \cup P_{s+r}$*.*

Then $(\star)$ *is moderate.*

Before we prove Theorem 6.23, we first show how it implies Theorem 6.8(ii).

Proof of Theorem 6.8(ii)*, assuming Theorem 6.23.* Let $C_1, \dots, C_\ell \subseteq [k]$ be the column equivalence classes of A. We distinguish two cases:

- If $\ell = 1$, then we have $m = \operatorname{rank}(A) \leq \ell = 1$, so we are in the situation with a single equation. Since we assumed $k \geq 3$, there is no column equivalence class of size 2, so it follows from Theorem 6.8(i) that $(\star)$ is moderate.

- Suppose that $\ell \geq 2$. Since A is non-degenerate, every column of A is non-zero. Hence, since the column equivalence classes of A sum to zero, every column equivalence class has size at least 2. For every $r \in [\ell]$, choose $j_r \in C_r$ arbitrary, and set $P_r := \{j_r\}$ and $P_{\ell+r} := C_r \setminus \{j_r\}$.

  We prove that the partition $[k] = P_1 \cup \dots \cup P_{2\ell}$ satisfies the properties from Theorem 6.23. Property (i) is met because each of the column equivalence classes sums to zero, and property (iii) is met by construction. To see that property (ii) is met, recall that $(\star)$ is irreducible, so it follows from Proposition 6.15 that every non-zero linear combination of the rows of A uses at least two different column equivalence classes. ■

Proof of Theorem 6.23. Let Γ_q be the constant from Theorem 6.17. We prove by induction on λ that, for every $\lambda \in [k]$, there is a constant $\beta_\lambda \geq 1$ such that every subset $S \subseteq \mathbb{F}_q^n$ of size at least $\beta_\lambda \cdot (\max(\Gamma_q, q^{\frac{k-1}{k}}))^n$ contains a solution $(\boldsymbol{x_1}, \dots, \boldsymbol{x_k}) \in S^k$ of $(\star)$ satisfying the following properties:

(a) the solution contains at least λ different vectors; that is, $|\{\boldsymbol{x_1}, \dots, \boldsymbol{x_k}\}| \geq \lambda$;

(b) for every column equivalence class of size 2 that sums to zero, the variables $\boldsymbol{x_{j_1}}, \boldsymbol{x_{j_2}}$ corresponding to that class are distinct.

Before proving the base case, we first show that the induction step from the proof of Theorem 6.8(i) carries through unchanged. This time, part (b) of the induction hypothesis replaces the assumption (i) from Theorem 6.8. To see that property (b) is automatically maintained by the proof of Theorem 6.8(i), recall that the induction step consists of choosing a column equivalence class C_t and replacing two variables from that class by other values, leaving the other classes unchanged. Since we started and ended with a solution of $(\star)$, the contribution of the variables $\{\boldsymbol{x_j} \mid j \in C_t\}$

[8] Note that we only look at $r \in \{1, \dots, s\}$, and we ignore all $r \in \{s+1, \dots, 2s\}$. This is because it follows from (i) that $\sum_{j \in P_r} b_j \neq 0$ if and only if $\sum_{j \in P_{s+r}} b_j \neq 0$. An equivalent statement is that $\sum_{j \in P_r} b_j \neq 0$ for at least four different values of $r \in [2s]$.

to $(\star)$ must have remained the same. Property (b) is equivalent to saying that the contribution of $\{\boldsymbol{x_j} \mid j \in C\}$ to $(\star)$ is non-zero for every column equivalence class C of size 2 that sums to zero, so this property is automatically maintained by the proof of Theorem 6.8(i).

It remains to prove the base case. Let $B = (b_{ir}) \in \mathbb{F}_q^{m \times s}$ be the matrix given by

$$b_{ir} := \sum_{j \in P_r} a_{ij} = -\sum_{j \in P_{s+r}} a_{ij}.$$

Suppose that $S \subseteq \mathbb{F}_q^n$ has size at least $q \cdot (\max(\Gamma_q, q^{\frac{k-1}{k}}))^n$. It follows from Lemma 6.22 that there are $(\boldsymbol{z_1}, \dots, \boldsymbol{z_s}), (\boldsymbol{z_{s+1}}, \dots, \boldsymbol{z_{2s}}) \in S^s$ such that, for all $(c_1, \dots, c_s) \in \mathbb{F}_q^s$, one has $c_1\boldsymbol{z_1} + \cdots + c_s\boldsymbol{z_s} = c_1\boldsymbol{z_{s+1}} + \cdots + c_s\boldsymbol{z_{2s}}$ if and only if $(c_1, \dots, c_s)$ is a linear combination of the rows of B. By assumption (ii), none of the standard unit vectors $\boldsymbol{e_1}, \dots, \boldsymbol{e_s} \in \mathbb{F}_q^s$ can be written as linear combination of the rows of B, so it follows that $\boldsymbol{z_r} \neq \boldsymbol{z_{s+r}}$ for all $r \in [s]$.

Since $[k] = P_1 \cup \cdots \cup P_{2s}$ is a partition, we may define $\boldsymbol{y_1}, \dots, \boldsymbol{y_k} \in \{\boldsymbol{z_1}, \dots, \boldsymbol{z_{2s}}\} \subseteq S$ in such a way that $\boldsymbol{y_j} = \boldsymbol{z_r}$ if and only if $j \in P_r$. Then for all $i \in [m]$ we have

$$\begin{aligned} a_{i1}\boldsymbol{y_1} + \cdots + a_{ik}\boldsymbol{y_k} &= \sum_{j \in P_1} a_{ij}\boldsymbol{z_1} + \cdots + \sum_{j \in P_{2s}} a_{ij}\boldsymbol{z_{2s}} \\ &= b_{i1}\boldsymbol{z_1} + \cdots + b_{is}\boldsymbol{z_s} - b_{i1}\boldsymbol{z_{s+1}} - \cdots - b_{is}\boldsymbol{z_{2s}} = 0, \end{aligned}$$

so $(\boldsymbol{y_1}, \dots, \boldsymbol{y_k}) \in S^k$ is a solution of $(\star)$. Clearly $|\{\boldsymbol{y_1}, \dots, \boldsymbol{y_k}\}| \geq 1$. Furthermore, by assumption (iii), for every column equivalence class $C = \{j_1, j_2\}$ of size 2 that sums to zero, there is some $r \in [s]$ such that $P_r = \{j_1\}$ and $P_{s+r} = \{j_2\}$, so it follows that $\boldsymbol{y_{j_1}} = \boldsymbol{z_r} \neq \boldsymbol{z_{s+r}} = \boldsymbol{y_{j_2}}$. ■

6.5 Preliminaries on temperate systems

We now shift our attention from moderate to temperate systems. We show that the problem of finding a generic solution is closely related to the problem of finding solutions of high dimension, and we show that we may once again restrict our attention to irreducible systems.

For an affine subspace $X \subseteq \mathbb{F}_q^n$ we let $\dim(X)$ denote the dimension of X. So $\dim(X)$ is the maximum number of affinely independent vectors in X minus one. For a set $S \subseteq \mathbb{F}_q^n$, we let $\mathrm{aff}(S)$ denote the affine hull of S.

Definition 6.24. For any given k-tuple $(\boldsymbol{x_1}, \dots, \boldsymbol{x_k}) \in (\mathbb{F}_q^n)^k$, let

$$\mathrm{Ann}_{\mathrm{bal}}(\boldsymbol{x_1}, \dots, \boldsymbol{x_k}) = \{(b_1, \dots, b_k) \in \mathbb{F}_q^k \mid b_1\boldsymbol{x_1} + \cdots + b_k\boldsymbol{x_k} = 0, \quad b_1 + \cdots + b_k = 0\}.$$

So the elements of $\mathrm{Ann}_{\mathrm{bal}}(\boldsymbol{x_1}, \dots, \boldsymbol{x_k})$ correspond to the balanced linear equations satisfied by $(\boldsymbol{x_1}, \dots, \boldsymbol{x_k})$.

Lemma 6.25. *For every $(\boldsymbol{x_1}, \dots, \boldsymbol{x_k}) \in (\mathbb{F}^n)^k$ we have*

$$\dim(\mathrm{aff}(\boldsymbol{x_1}, \dots, \boldsymbol{x_k})) + \dim(\mathrm{Ann}_{\mathrm{bal}}(\boldsymbol{x_1}, \dots, \boldsymbol{x_k})) = k - 1.$$

Proof. Let $A \in \mathbb{F}^{(n+1)\times k}$ be the matrix

$$A = \begin{pmatrix} 1 & \cdots & 1 \\ | & & | \\ \boldsymbol{x_1} & \cdots & \boldsymbol{x_k} \\ | & & | \end{pmatrix}.$$

For $I \subseteq [k]$ the vectors $\boldsymbol{x_i}$, $i \in I$ are affinely independent if and only if the columns of A indexed by I are linearly independent. So $\operatorname{rank}(A) = \dim(\operatorname{aff}(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k})) + 1$.

Evidently, $\ker(A)$ is precisely $\operatorname{Ann}_{\text{bal}}(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k})$, so the result follows from the rank-nullity theorem. ■

Corollary 6.26. *Let $(\star)$ be a balanced linear system of rank m, with coefficient matrix $A \in \mathbb{F}_q^{m\times k}$, and let $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k})$ be a solution of $(\star)$. Then $\dim(\operatorname{aff}(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k})) \leq k - m - 1$, with equality if and only if $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k})$ is a generic solution of $(\star)$.*

Proof. Since $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k})$ is a solution of $(\star)$, the row space of A is contained in $\operatorname{Ann}_{\text{bal}}(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k})$. Therefore we have $m = \operatorname{rank}(A) \leq \dim(\operatorname{Ann}_{\text{bal}}(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}))$, so it follows from Lemma 6.25 that

$$\dim(\operatorname{aff}(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k})) = k - 1 - \dim(\operatorname{Ann}_{\text{bal}}(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k})) \leq k - 1 - m.$$

Clearly we have equality if and only if the row space of A is equal to $\operatorname{Ann}_{\text{bal}}(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k})$, which is equivalent to saying that all balanced linear equations satisfied by $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k})$ are linear combinations of the equations in $(\star)$. ■

Proposition 6.27. *Suppose that $(\star)$ is equivalent to a linear system $(\star')$ whose coefficient matrix A' can be written as*

$$A' = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$$

for some $A_1 \in \mathbb{F}_q^{m_1\times k_1}$ and $A_2 \in \mathbb{F}_q^{m_2\times k_2}$ with $m_1, m_2, k_1, k_2 \neq 0$. Then $(\star)$ is temperate if and only if the systems given by A_1 and A_2 are temperate.

Proof. If $(\star')$ is temperate, then it is easy to see that the same holds for the systems given by A_1 and A_2.

Suppose that for $i = 1, 2$ the system given by A_i is temperate, with constants $\beta_i, \gamma_i > 0$, where $\gamma_i < q$. Let γ satisfy $\max(\gamma_1, \gamma_2) < \gamma < q$, and choose β such that

$$\beta q^{\gamma n} \geq \max(qn \cdot \beta_1 q^{\gamma_1 n},\ nq^{k_1} \cdot \beta_2 q^{\gamma_2 n}) \qquad \text{for all } n \in \mathbb{N}_1.$$

Let $S \subseteq \mathbb{F}_q^n$ have size $|S| \geq \beta q^{\gamma n}$. For $i \in [n]$ and $\alpha \in \mathbb{F}_q$, write $S(i, \alpha) := \{\boldsymbol{x} \in S \mid x_i = \alpha\}$. We claim that there exist $i \in [n]$ and distinct $\alpha', \alpha'' \in \mathbb{F}_q$ such that $|S(i, \alpha')|, |S(i, \alpha'')| \geq \frac{|S|}{qn}$. For each coordinate $i \in [n]$, let $\alpha_i \in \arg\max_{\alpha\in\mathbb{F}_q} |S(i, \alpha)|$ be a most popular value. Then $S \setminus \{(\alpha_1, \ldots, \alpha_n)\} = \cup_{i\in[n]}(S \setminus S(i, \alpha_i))$. So we can choose $i \in [n]$ such that $|S \setminus S(i, \alpha_i)| \geq \frac{|S|-1}{n}$. Then there is an $\alpha'' \neq \alpha_i$ such that $S(i, \alpha'') \geq \frac{|S|-1}{n(q-1)} \geq \frac{|S|}{qn}$. Taking $\alpha' = \alpha_i$ proves the claim.

Without loss of generality, we will assume that we can take $i = 1$ in the claim. We denote $S_1 = S(1, \alpha')$ and $S_2 = S(1, \alpha'')$. Since $|S_1| \geq \beta_1 q^{\gamma_1 n}$, there exists a generic solution $\vec{\boldsymbol{y}} = (\boldsymbol{y_1}, \ldots, \boldsymbol{y_{k_1}}) \in (S_1)^{k_1}$ to the linear system given by A_1. We can take $I \subseteq [n]$ with $|I| \leq k_1 - 1$ such that for all $\boldsymbol{b} = (b_1, \ldots, b_{k_1}) \in \mathbb{F}_q^{k_1}$ with $b_1 + \cdots + b_{k_1} = 0$ we have:

$$\forall i \in I : (b_1 \boldsymbol{y_1} + \cdots + b_{k_1} \boldsymbol{y_{k_1}})_i = 0 \implies b_1 \boldsymbol{y_1} + \cdots + b_{k_1} \boldsymbol{y_{k_1}} = 0.$$

Indeed, if $M \in \mathbb{F}_q^{n \times k_1}$ is the matrix with columns $\boldsymbol{y_1}, \ldots, \boldsymbol{y_{k_1}}$, then we can take $I \subseteq [n]$ of size $|I| \leq k_1 - 1$ such that the rows of M are contained in the span of the rows indexed by I and the row vector $(1, \ldots, 1)$. Since $y_{11} = \cdots = y_{k_1 1}$ we may assume that $1 \notin I$.

As $\vec{\boldsymbol{y}}$ is a generic solution to the system given by A_1, we obtain

$$\forall i \in I : (b_1 \boldsymbol{y_1} + \cdots + b_{k_1} \boldsymbol{y_{k_1}})_i = 0 \implies \boldsymbol{b} \in \text{rowspace}(A_1). \tag{6.28}$$

We can take $\alpha_i \in \mathbb{F}_q$ for each $i \in I$ such that $T = \{\boldsymbol{x} \in S_2 \mid x_i = \alpha_i \text{ for all } i \in I\}$ has size $|T| \geq |S_2| \cdot q^{1-k_1} \geq \beta_2 q^{\gamma_2 n}$.

It follows that there exists a generic solution $\vec{\boldsymbol{z}} \in T^{k_2}$ to the system given by A_2. Now $\vec{\boldsymbol{x}} = (\vec{\boldsymbol{y}}, \vec{\boldsymbol{z}})$ is a generic solution to $(\star')$. Indeed, let $\boldsymbol{b} = (b_1, \ldots, b_k) \in \text{Ann}_{\text{bal}}(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k})$. It suffices to show that $\boldsymbol{b} \in \text{rowspace}(A')$. Looking at the first coordinate and using that $b_1 + \cdots + b_k = 0$, we see that

$$0 = (b_1 + \cdots + b_{k_1})\alpha' + (b_{k_1+1} + \cdots + b_k)\alpha'' = (b_1 + \cdots + b_{k_1})(\alpha' - \alpha'').$$

Since $\alpha' \neq \alpha''$, we find that $b_1 + \cdots + b_{k_1} = 0 = b_{k_1+1} + \cdots + b_k$. Since $\vec{\boldsymbol{z}} \in T^{k_2}$ it follows that

$$(b_1 \boldsymbol{y_1} + \cdots + b_{k_1} \boldsymbol{y_{k_1}})_i = (b_1 \boldsymbol{x_1} + \cdots + b_k \boldsymbol{x_k})_i = 0 \ (\forall i \in I).$$

It now follows from (6.28) that $(b_1, \ldots, b_{k_1}) \in \text{rowspace}(A_1)$. So after modifying $\boldsymbol{b}$ by an element of $\text{rowspace}(A')$, we may assume that $b_1, \ldots, b_{k_1} = 0$. Hence the fact that $\boldsymbol{b} \in \text{Ann}_{\text{bal}}(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k})$ implies that $b_{k_1+1} \boldsymbol{z_1} + \cdots + b_k \boldsymbol{z_{k_2}} = 0$. Since $\vec{\boldsymbol{z}}$ is generic, we conclude that $(b_{k_1+1}, \ldots, b_k) \in \text{rowspace}(A_2)$. Hence, $\boldsymbol{b} \in \text{rowspace}(A')$. ■

6.6 Proof of Theorem 6.9 and Theorem 6.10

In this section, we develop the multiple replacement trick (Corollary 6.30) and use it (in combination with Lemma 6.22) to prove Theorem 6.9 and Theorem 6.10.

We start with a many-solutions version of Lemma 6.18.

Lemma 6.29. *Let q be a prime power, let $N_0 = (\Gamma_q)^n$, where Γ_q is as in Theorem 6.17, and let $t \in \mathbb{N}_1$. Let $\boldsymbol{x_1}, \ldots, \boldsymbol{x_L} \in \mathbb{F}_q^n$ be distinct, let $\boldsymbol{y_1}, \ldots, \boldsymbol{y_L} \in \mathbb{F}_q^n$ be distinct, and let $\alpha, \beta \in \mathbb{F}_q \setminus \{0\}$. If $L \geq 4tN_0$, then there exists an $i \in [L]$ such that*

$$\left|\left\{(i', i'') \in ([L] \setminus \{i\})^2 \mid \alpha \boldsymbol{x_{i'}} + \beta \boldsymbol{y_{i''}} = \alpha \boldsymbol{x_i} + \beta \boldsymbol{y_i}\right\}\right| \geq t.$$

Proof. Write

$$T := \{(i, i', i'') \in [L]^3 \mid \alpha \boldsymbol{x_{i'}} + \beta \boldsymbol{y_{i''}} = \alpha \boldsymbol{x_i} + \beta \boldsymbol{y_i} \text{ and } i \neq i', i''\}.$$

By Lemma 6.18, the set $T \cap J^3$ is nonempty for all $J \subseteq [L]$ with $|J| \geq N_0$. We claim that $|T \cap J^3| \geq |J| - N_0$ for all $J \subseteq [L]$. Indeed, suppose that $|T \cap J^3| < |J| - N_0$; then we could delete fewer than $|J| - N_0$ elements from J to obtain a set J' of size $|J'| > N_0$ such that $T \cap (J')^3$ is empty: a contradiction. So $|T \cap J^3| - |J| + N_0 \geq 0$ for all $J \subseteq [L]$.

Let J be the random subset of $[L]$ obtained by independently taking each element of $[L]$ with probability $\frac{1}{2t}$. We have $\mathbb{E}[|J|] = \frac{L}{2t}$ and $\mathbb{E}[|T \cap J^3|] \leq \frac{|T|}{(2t)^2}$ since $|\{i, i', i''\}| \geq 2$ for all $(i, i', i'') \in T$. From $\mathbb{E}[|T \cap J^3| - |J| + N_0] \geq 0$ we obtain $\frac{|T|}{4t^2} \geq \frac{L}{2t} - N_0$, and therefore $\frac{|T|}{L} \geq 2t - \frac{4t^2 N_0}{L} \geq t$. Hence, by the pigeonhole principle, there is an $i \in [L]$ such that $|\{(i', i'') \in [L]^2 \mid (i, i', i'') \in T\}| \geq t$, as required. ∎

Recall that two solutions $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k})$ and $(\boldsymbol{y_1}, \ldots, \boldsymbol{y_k})$ are said to be disjoint if $\{\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}\} \cap \{\boldsymbol{y_1}, \ldots, \boldsymbol{y_k}\} = \varnothing$. We obtain a corollary analogous to Corollary 6.21.

Corollary 6.30 ('Multiple replacement trick')**.** *Let $\{(\boldsymbol{x_1^{(i)}}, \ldots, \boldsymbol{x_k^{(i)}})\}_{i=1}^{L}$ be a list of pairwise disjoint solutions of $(\star)$, and suppose that j_1 and j_2 are distinct indices from the same column equivalence class. Suppose that $L \geq 4t \cdot (\Gamma_q)^n$. Then there exist an $i \in [L]$ and t distinct pairs $(i'_s, i''_s) \in ([L] \setminus \{i\})^2$, $s \in [t]$, such that $(\boldsymbol{y_1^{(s)}}, \ldots, \boldsymbol{y_k^{(s)}}) \in (\mathbb{F}_q^n)^k$ given by*

$$\boldsymbol{y_j^{(s)}} = \begin{cases} \boldsymbol{x_j^{(i)}}, & \text{if } j \neq j_1, j_2; \\ \boldsymbol{x_j^{(i'_s)}}, & \text{if } j = j_1; \\ \boldsymbol{x_j^{(i''_s)}}, & \text{if } j = j_2; \end{cases}$$

is also a solution of $(\star)$ for all $s \in [t]$.

Proof. Since the j_1-th and j_2-th column of $(\star)$ are nonzero multiples of one another, we may choose a vector $\boldsymbol{v} \in \mathbb{F}_q^m$ and constants $\alpha, \beta \neq 0$ such that the j_1-th column is equal to $\alpha \boldsymbol{v}$ and the j_2-th column is equal to $\beta \boldsymbol{v}$.

By assumption, the vectors $\boldsymbol{x_{j_1}^{(1)}}, \ldots, \boldsymbol{x_{j_1}^{(L)}}$ are pairwise distinct, and likewise the vectors $\boldsymbol{x_{j_2}^{(1)}}, \ldots, \boldsymbol{x_{j_2}^{(L)}}$ are pairwise distinct, so it follows from Lemma 6.29 that there exist $i \in [L]$ and t distinct pairs $(i'_s, i''_s) \in ([L] \setminus \{i\})^2$, $s \in [t]$, with $\alpha \boldsymbol{x_{j_1}^{(i)}} + \beta \boldsymbol{x_{j_2}^{(i)}} = \alpha \boldsymbol{x_{j_1}^{(i'_s)}} + \beta \boldsymbol{x_{j_2}^{(i''_s)}}$. Hence, the total contribution of $\boldsymbol{x_{j_1}^{(i)}}$ and $\boldsymbol{x_{j_2}^{(i)}}$ to the equations of $(\star)$ is the same as the contribution of $\boldsymbol{x_{j_1}^{(i'_s)}}$ and $\boldsymbol{x_{j_2}^{(i''_s)}}$. Since $(\boldsymbol{x_1^{(i)}}, \ldots, \boldsymbol{x_k^{(i)}})$ is a solution of $(\star)$, so is $(\boldsymbol{y_1^{(s)}}, \ldots, \boldsymbol{y_k^{(s)}})$. ∎

Definition 6.31. Let $A \in \mathbb{F}_q^{m \times k}$ be a matrix and let $j_1, j_2 \in [k]$ be distinct elements in the same column equivalence class of A. We say that $(b_1, \ldots, b_k) \in \mathbb{F}_q^k$ *breaks* the pair $\{j_1, j_2\}$ if after adding the row $(b_1, \ldots, b_k)$ to A, the columns indexed by j_1 and j_2 are no longer scalar multiples of one another.

Lemma 6.32. *Let $(\star)$, A, m, k and ℓ be as in Situation 6.7, let $j_1, j_2 \in [k]$ be distinct indices in the same column equivalence class, and let $\{(\boldsymbol{x}_1^{(i)}, \ldots, \boldsymbol{x}_k^{(i)})\}_{i=1}^{L}$ be a list of pairwise disjoint solutions to $(\star)$. If $L \geq 4q^k(\Gamma_q)^n$, then there exist an $i \in [L]$ and a solution $(\boldsymbol{y_1}, \ldots, \boldsymbol{y_k})$ to $(\star)$ such that:*

(i) $\boldsymbol{y_j} = \boldsymbol{x_j^{(i)}}$ *for all* $j \neq j_1, j_2$ *and* $\boldsymbol{y_j} \in \{\boldsymbol{x_j^{(1)}}, \ldots, \boldsymbol{x_j^{(L)}}\}$ *for* $j \in \{j_1, j_2\}$;

(ii) $\mathrm{Ann}_{\mathrm{bal}}(\boldsymbol{y_1}, \ldots, \boldsymbol{y_k}) \subseteq \mathrm{Ann}_{\mathrm{bal}}(\boldsymbol{x_1^{(i)}}, \ldots, \boldsymbol{x_k^{(i)}})$;

(iii) *no* $\boldsymbol{b} \in \mathrm{Ann}_{\mathrm{bal}}(\boldsymbol{y_1}, \ldots, \boldsymbol{y_k})$ *breaks the pair* (j_1, j_2).

Proof. By Corollary 6.30, we may choose $i \in [L]$ and a sequence $\{(i'_s, i''_s)\}_{s=1}^{q^k}$ of q^k pairwise distinct pairs $(i'_s, i''_s) \in ([L] \setminus \{i\})^2$ such that, for all $s \in [q^k]$, the k-tuple $(\boldsymbol{z_1^{(s)}}, \ldots, \boldsymbol{z_k^{(s)}}) \in S^k$ defined by

$$\boldsymbol{z_j^{(s)}} = \begin{cases} \boldsymbol{x_j^{(i)}} & \text{if } j \in [k] \setminus \{j_1, j_2\} \\ \boldsymbol{x_j^{(i'_s)}} & \text{if } j = j_1 \\ \boldsymbol{x_j^{(i''_s)}} & \text{if } j = j_2 \end{cases}$$

is a solution to $(\star)$.

If $\boldsymbol{b} = (b_1, \ldots, b_k)$ breaks the pair (j_1, j_2), then the contributions $b_{j_1}\boldsymbol{z_{j_1}^{(s)}} + b_{j_2}\boldsymbol{z_{j_2}^{(s)}}$ for $s \in [q^k]$ are pairwise distinct. Therefore we can have $\boldsymbol{b} \in \mathrm{Ann}_{\mathrm{bal}}(\boldsymbol{z_1^{(s)}}, \ldots, \boldsymbol{x_k^{(s)}})$ for at most one value of s. Since the number of $\boldsymbol{b} \in \mathbb{F}_q^k$ with $b_1 + \cdots + b_k = 0$ is less than q^k, we may choose $s_0 \in [q^k]$ such that no $\boldsymbol{b} \in \mathrm{Ann}_{\mathrm{bal}}(\boldsymbol{z_1^{(s_0)}}, \ldots, \boldsymbol{z_k^{(s_0)}})$ breaks the pair (j_1, j_2).

Set $y := \boldsymbol{z^{(s_0)}}$. Then (i) and (iii) are met. To prove (ii), let $\boldsymbol{b} \in \mathrm{Ann}_{\mathrm{bal}}(\boldsymbol{y_1}, \ldots, \boldsymbol{y_k})$ be given. Since $\boldsymbol{b}$ does not break the pair (j_1, j_2), we have $b_{j_1}\boldsymbol{z_{j_1}^{(s_0)}} + b_{j_2}\boldsymbol{z_{j_2}^{(s_0)}} = b_{j_1}\boldsymbol{x_{j_1}^{(i)}} + b_{j_2}\boldsymbol{x_{j_2}^{(i)}}$, and therefore $\boldsymbol{b} \in \mathrm{Ann}_{\mathrm{bal}}(\boldsymbol{x_1^{(i)}}, \ldots, \boldsymbol{x_k^{(i)}})$, as desired. ∎

Lemma 6.33. *Let $(\star)$, A, m, k and ℓ be as in Situation 6.7. Let $S \subseteq \mathbb{F}_q^n$ have size $|S| \geq q^{1+\frac{\ell-1}{\ell}n}$. Assume that at least one of the following two conditions holds:*

(i) $\ell = m + 1$;

(ii) *every column equivalence class sums to zero.*

Then there exists a solution $\vec{\boldsymbol{x}} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}) \in S^k$ to $(\star)$ with the following property:

$$\begin{gathered} \textit{If } \boldsymbol{b} \in \mathrm{Ann}_{\mathrm{bal}}(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}) \textit{ preserves the column} \\ \textit{equivalence classes of } (\star), \textit{ then } \boldsymbol{b} \in \mathrm{rowspace}(A). \end{gathered} \tag{6.34}$$

Proof. Let $[k] = C_1 \cup \cdots \cup C_\ell$ be the partition of $[k]$ into column equivalence classes.

We first consider the case that condition (i) holds. Let $\vec{\boldsymbol{x}} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_k})$ be any solution to $(\star)$. Suppose that $\vec{\boldsymbol{x}}$ satisfies a balanced equation $b_1\boldsymbol{x_1} + \cdots + b_k\boldsymbol{x_k} = 0$ that preserves the column equivalence classes of $(\star)$, but $(b_1, \ldots, b_k)$ is not a linear

combination of the rows of A. Let A' be the $(m+1)\times k$ matrix obtained by adding the row $(b_1,\dots,b_k)$ to A. Then $\operatorname{rank}(A') = m+1 = \ell$. For $t\in[\ell]$ let $\boldsymbol{\sigma_t}\in\mathbb{F}_q^{m+1}$ be the sum of the columns of A' in class t. Since the column rank of A' is ℓ, it follows that if we take one index from each column equivalence class, the corresponding ℓ columns are linearly independent. Let $I=\{t\in[\ell]\mid \boldsymbol{\sigma_t}\neq 0\}$. Then the $\boldsymbol{\sigma_t}$, $t\in I$ are linearly independent and $\sum_{t\in I}\boldsymbol{\sigma_t}=\sum_{t\in[\ell]}\boldsymbol{\sigma_t}=0$. It follows that $I=\varnothing$. So all column equivalence classes of A' (and hence of A) sum to zero, and we are in case (ii).

We now consider the case that condition (ii) holds. Since A has no zero columns, every column equivalence class has size at least 2. For $t\in[\ell]$ let $j_t\in C_t$. Let $A'=(a'_{it})\in\mathbb{F}_q^{m\times\ell}$ be the submatrix of A induced by columns $j_1,\dots,j_t$. Consider the system

$$\sum_{t=1}^{\ell} a'_{it}\boldsymbol{y_t}=0 \text{ for all } i\in[m].$$

Since $|S|\geq q^{1+\frac{\ell-1}{\ell}n}$, it follows by Lemma 6.22 that there are $(\boldsymbol{y_1},\dots,\boldsymbol{y_\ell})$ and $(\boldsymbol{z_1},\dots,\boldsymbol{z_\ell})$ in S^ℓ such that for all $(b_1,\dots,b_\ell)\in\mathbb{F}_q^\ell$ one has $b_1(\boldsymbol{y_1}-\boldsymbol{z_1})+\cdots+b_\ell(\boldsymbol{y_\ell}-\boldsymbol{z_\ell})=0$ if and only if $(b_1,\dots,b_\ell)$ is a linear combination of the rows of A'. Define $(\boldsymbol{x_1},\dots,\boldsymbol{x_k})\in S^k$ by setting (for $t\in[\ell]$ and $j\in C_t$)

$$\boldsymbol{x_j}=\begin{cases}\boldsymbol{y_t} & \text{if } j=j_t,\\ \boldsymbol{z_t} & \text{otherwise.}\end{cases}$$

Note that for any $b_{j_1},\dots,b_{j_\ell}\in\mathbb{F}_q$ there are unique $b_j\in\mathbb{F}_q$, $j\in[k]\setminus\{j_1,\dots,j_\ell\}$ such that $b_1+\cdots+b_k=0$ and $(b_1,\dots,b_k)$ preserves the column equivalence classes of A. Moreover, we have $b_{j_1}(\boldsymbol{y_1}-\boldsymbol{z_1})+\cdots+b_{j_\ell}(\boldsymbol{y_\ell}-\boldsymbol{z_\ell})=0$ if and only if $b_1\boldsymbol{x_1}+\cdots+b_k\boldsymbol{x_k}=0$. It follows that $(\boldsymbol{x_1},\dots,\boldsymbol{x_k})$ satisfies the requirements. ■

We are now ready to prove Theorem 6.9 and Theorem 6.10.

Proof of Theorem 6.9. Let Γ_q be the constant from Theorem 6.17. For every $t\in\mathbb{N}_0$, we define

$$N_t := q^{1+\frac{\ell-1}{\ell}n}+t\cdot(4kq^k(\Gamma_q)^n).$$

Let $[k]=C_1\cup\cdots\cup C_\ell$ be the partition of $[k]$ into column equivalence classes of A. We will prove by induction on $|P|$ that, for every set $P\subseteq\binom{C_1}{2}\cup\cdots\cup\binom{C_\ell}{2}$ of equivalent pairs and for every set $S\subseteq\mathbb{F}_q^n$ of size $|S|\geq N_{|P|}$, the system $(\star)$ has a solution $\vec{\boldsymbol{x}}=(\boldsymbol{x_1},\dots,\boldsymbol{x_k})\in S^k$ that satisfies (6.34) and such that no $(b_1,\dots,b_k)\in\operatorname{Ann}_{\mathrm{bal}}(\boldsymbol{x_1},\dots,\boldsymbol{x_k})$ breaks a pair in P.

- For $|P|=0$, the claim follows directly from Lemma 6.33.
- Assume that $|P|\geq 1$ and that the claim holds for sets of fewer than $|P|$ pairs. Fix some $\{j_1,j_2\}\in P$, and write $L=4q^k(\Gamma_q)^n$. Since $|S|\geq N_{|P|}\geq kL+N_{|P|-1}$, there exist disjoint solutions $\vec{\boldsymbol{x}}^{\,(\mathbf{1})},\dots,\vec{\boldsymbol{x}}^{\,(\boldsymbol{L})}\in S^k$ to $(\star)$ that satisfy the desired property for the list $P\setminus\{\{j_1,j_2\}\}$. Using Lemma 6.32, we obtain a solution $\vec{\boldsymbol{x}}=(\boldsymbol{x_1},\dots,\boldsymbol{x_k})\in S^k$ to $(\star)$ with $\operatorname{Ann}_{\mathrm{bal}}(\boldsymbol{x_1},\dots,\boldsymbol{x_k})\subseteq\operatorname{Ann}_{\mathrm{bal}}(\boldsymbol{x_1^{(i)}},\dots,\boldsymbol{x_k^{(i)}})$ for some $i\in[L]$ and such that no $\boldsymbol{b}\in\operatorname{Ann}_{\mathrm{bal}}(\boldsymbol{x_1},\dots,\boldsymbol{x_k})$ breaks the pair (j_1,j_2).

Since $\mathrm{Ann}_{\mathrm{bal}}(\boldsymbol{x_1}, \dots, \boldsymbol{x_k}) \subseteq \mathrm{Ann}_{\mathrm{bal}}(\boldsymbol{x_1^{(i)}}, \dots, \boldsymbol{x_k^{(i)}})$, no $\boldsymbol{b} \in \mathrm{Ann}_{\mathrm{bal}}(\boldsymbol{x_1}, \dots, \boldsymbol{x_k})$ breaks a pair in $P \setminus \{(j_1, j_2)\}$ and $\vec{\boldsymbol{x}}$ satisfies (6.34).

Letting $P = \binom{C_1}{2} \cup \dots \cup \binom{C_\ell}{2}$ completes the proof. ■

Proof of Theorem 6.10. If all column equivalence classes sum to zero, the result follows directly from Theorem 6.9(ii). Assume therefore that not all column equivalence classes sum to zero. Let Γ_q be the constant from Theorem 6.17. For every $t \in \mathbb{N}_0$ we define

$$N_t := t \cdot (4kq^k(\Gamma_q)^n).$$

Let $S \subseteq \mathbb{F}_q^n$ have size $|S| \geq N_{k_2}$. By the same argument as in the proof of Theorem 6.9, we have a solution $\vec{\boldsymbol{x}} = (\boldsymbol{x_1}, \dots, \boldsymbol{x_k}) \in S^k$ to $(\star)$ such that no $\boldsymbol{b} \in \mathrm{Ann}_{\mathrm{bal}}(\boldsymbol{x_1}, \dots, \boldsymbol{x_k})$ breaks a pair from the same column equivalence class. In other words, $\boldsymbol{b}$ preserves the column equivalence classes, so this proves part (i).

For part (ii), observe that $\mathrm{Ann}_{\mathrm{bal}}(\boldsymbol{x_1}, \dots, \boldsymbol{x_k})$ does not contain all balanced linear equations that preserve the column equivalence classes, for otherwise every column equivalence class must sum to zero, contrary to our assumption. So we have $\dim(\mathrm{Ann}_{\mathrm{bal}}(\boldsymbol{x_1}, \dots, \boldsymbol{x_k})) \leq \ell - 1$, and therefore $\dim(\mathrm{aff}(\boldsymbol{x_1}, \dots, \boldsymbol{x_k})) \geq k - \ell$, by Lemma 6.25. ■

Remark 6.35. We compare the rank of the solution $(\boldsymbol{x_1}, \dots, \boldsymbol{x_k})$ in Theorem 6.10 to the rank given by Theorem 6.5. Suppose we are in Situation 6.7, and set $r = k - 2m + 1$. Then $k \geq 2m - 1 + r$, so it follows from Theorem 6.5 that we can find a solution with $\dim(\mathrm{span}(\boldsymbol{x_1}, \dots, \boldsymbol{x_k})) \geq r$, and therefore $\dim(\mathrm{aff}(\boldsymbol{x_1}, \dots, \boldsymbol{x_k})) \geq r - 1 = k - 2m$.

So how do these two compare? If $\ell = 1$, then we must have $m = 1$ (because we assume that the rows of A are linearly independent), so in this case the rank from Theorem 6.10 and Theorem 6.5 agree. If $\ell \geq 2$, then we see that Theorem 6.10 improves upon Theorem 6.5 whenever $m > \frac{\ell}{2}$. Then again, Theorem 6.10 only applies to a smaller class of linear systems.

6.7 Examples and applications

We conclude this chapter by looking at a few examples of type (RC) linear systems, to highlight the applications and limitations of the results from this chapter. First we will look at an application to sumsets in $\mathbb{F}_q^n$. We show that our results can be used to find non-trivial solutions of an arbitrary linear system in the difference set $S - S$, but not in the sumset $S + S$. After that, we will look at the systems studied by Mimura and Tokushige [MT19a, MT19b, MT20]. We show that our techniques furnish alternative proofs that those systems are moderate, and in many cases we strengthen this to show that the system is also temperate.

Applications to sum and difference sets

Since this chapter studies linear systems with repeated columns, one obvious question is to which extent our results can be applied to the problem of finding solutions to a system of linear equations in sum and difference sets. Throughout this section, let

$\mathbb{F}_q$ be a finite field of characteristic p, and let $c_1, \dots, c_l \in \mathbb{F}_q \setminus \{0\}$. We consider the affinely independent sumset (or AIR-sumset)

$$T := c_1 \cdot S \dot{+}_{\text{aff}} \cdots \dot{+}_{\text{aff}} c_l \cdot S = \{c_1 \boldsymbol{x_1} + \cdots + c_l \boldsymbol{x_l} \mid \boldsymbol{x_1}, \dots, \boldsymbol{x_l} \in S \text{ affinely independent}\}.$$

If $c_1 + \cdots + c_l = 0$, then Corollary 6.11 states that T contains generic solutions to every linear system $(\star)$, provided that S is sufficiently large. We now prove this statement.

Proof of Corollary 6.11. Let $A = (a_{ij}) \in \mathbb{F}_q^{m \times k}$ be the coefficient matrix of the system $(\star)$. (Recall from the statement of Corollary 6.11 that A may be arbitrary.) Let $A' = (a'_{ij}) \in \mathbb{F}_q^{m \times lk}$ be the $m \times lk$ matrix

$$A' = \left[c_1 A \mid c_2 A \mid \cdots \mid c_l A\right],$$

and let $(\star')$ be the corresponding linear system. Every column equivalence class of $(\star')$ is the union of sets of the form $\{j, j+k, \dots, j+(l-1)k\}$ (for some $j \in [k]$), so $(\star')$ is of type (RC). Furthermore, the column equivalence classes sum to zero, because $c_1 + \cdots + c_l = 0$. Hence it follows from Theorem 6.9(ii) and Proposition 6.27 that $(\star')$ is temperate. Therefore there are constants $\beta, \gamma \geq 1$ with $\gamma < q$ such that every set $S \subseteq \mathbb{F}_q^n$ with $|S| \geq \beta \cdot \gamma^n$ contains a generic solution of $(\star')$. Choose such a generic solution $(\boldsymbol{x_1}, \dots, \boldsymbol{x_{lk}}) \in S^{lk}$, and define $\boldsymbol{y_1}, \dots, \boldsymbol{y_k} \in c_1 \cdot S + \cdots + c_l \cdot S$ by

$$\boldsymbol{y_j} := c_1 \boldsymbol{x_j} + c_2 \boldsymbol{x_{j+k}} + \cdots + c_l \boldsymbol{x_{j+(l-1)k}}.$$

Clearly $(\boldsymbol{y_1}, \dots, \boldsymbol{y_k})$ is a solution of the linear system $(\star)$. We show that $(\boldsymbol{y_1}, \dots, \boldsymbol{y_k})$ is linearly generic and that $\boldsymbol{y_1}, \dots, \boldsymbol{y_k} \in (c_1 \cdot S \dot{+}_{\text{aff}} \cdots \dot{+}_{\text{aff}} c_l \cdot S) \cup \{0\}$.

First, let $\boldsymbol{b} = (b_1, \dots, b_k) \in \mathbb{F}_q^k$ be such that $b_1 \boldsymbol{y_1} + \cdots + b_k \boldsymbol{y_k} = 0$. Then $(\boldsymbol{x_1}, \dots, \boldsymbol{x_{lk}})$ belongs to the kernel of the $1 \times lk$ matrix

$$B' = \left[c_1 \boldsymbol{b} \mid c_2 \boldsymbol{b} \mid \cdots \mid c_l \boldsymbol{b}\right].$$

Since $c_1 + \cdots + c_l = 0$, the entries of B' sum to 0, so B' represents a balanced linear equation satisfied by $(\boldsymbol{x_1}, \dots, \boldsymbol{x_{lk}})$. Since $(\boldsymbol{x_1}, \dots, \boldsymbol{x_{lk}})$ is a generic solution of $(\star')$, it follows that B' is a linear combination of the rows of A'. Equivalently, $\boldsymbol{b}$ is a linear combination of the rows of A. This shows that $(\boldsymbol{y_1}, \dots, \boldsymbol{y_k})$ is linearly generic.

To complete the proof, it suffices to show that $\boldsymbol{y_j} = 0$ whenever the vectors $\boldsymbol{x_j}, \boldsymbol{x_{j+k}}, \dots, \boldsymbol{x_{j+(l-1)k}}$ are affinely dependent, for every $j \in [k]$. To that end, suppose that $\boldsymbol{x_j}, \boldsymbol{x_{j+k}}, \dots, \boldsymbol{x_{j+(l-1)k}}$ are affinely dependent. Then there is some $\boldsymbol{b} = (b_1, \dots, b_l) \in \mathbb{F}_q^l \setminus \{0\}$ with $b_1 + \cdots + b_l = 0$ and

$$b_1 \boldsymbol{x_j} + b_2 \boldsymbol{x_{j+k}} + \cdots + b_l \boldsymbol{x_{j+(l-1)k}} = 0. \qquad (\boldsymbol{b'})$$

Since $(\boldsymbol{x_1}, \dots, \boldsymbol{x_{lk}})$ is generic, the balanced linear equation $(\boldsymbol{b'})$ is a linear combination of the equations in $(\star')$. By choosing some $r \in [l]$ such that $b_r \neq 0$ and restricting our attention to the variables $\boldsymbol{x_{(r-1)k+1}}, \dots, \boldsymbol{x_{rk}}$ (i.e. the r-th block in the block matrix representation of A'), we see that the equation $\boldsymbol{y_j} = 0$ is a linear combination of the equations in $(\star)$. ∎

Corollary 6.12 can be deduced from Corollary 6.11 by letting $(\star)$ be the linear system that encodes a k-term arithmetic progression and setting $l = 2$ and $(c_1, c_2) = (1, -1)$. We show that Corollary 6.12 does not depend on the full strength of Corollary 6.11, as it follows immediately from Lemma 6.22.

Proof of Corollary 6.12. Let $(\star)$ be a linear system which encodes a k-term arithmetic progression, for instance the system given by the matrix

$$A = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & -2 & 1 \end{pmatrix} \in \mathbb{F}_p^{(k-2)\times k}.$$

Let $S \subseteq \mathbb{F}_p^n$ with $|S| \geq p^{1+(1-\frac{1}{k})n}$. By Lemma 6.22, there are $(\boldsymbol{x_1}, \dots, \boldsymbol{x_k}), (\boldsymbol{y_1}, \dots, \boldsymbol{y_k}) \in S^k$ such that $(\boldsymbol{x_1} - \boldsymbol{y_1}, \dots, \boldsymbol{x_k} - \boldsymbol{y_k})$ is a linearly generic solution of $(\star)$.

Since the standard basis vectors $\boldsymbol{e_1}, \dots, \boldsymbol{e_k} \in \mathbb{F}_q^k$ cannot be written as linear combinations of the rows of A,[9] we have $\boldsymbol{x_j} - \boldsymbol{y_j} \neq 0$ for all $j \in [k]$. Likewise, since the vectors $\boldsymbol{e_j} - \boldsymbol{e_{j'}}$ $(j \neq j')$ cannot be written as linear combinations of the rows of A,[9] we have $\boldsymbol{x_j} - \boldsymbol{y_j} \neq \boldsymbol{x_{j'}} - \boldsymbol{y_{j'}}$ whenever $j \neq j'$. It follows that $(\boldsymbol{x_1} - \boldsymbol{y_1}, \dots, \boldsymbol{x_k} - \boldsymbol{y_k})$ is a non-trivial k-AP in $(S - S) \setminus \{0\}$. ∎

Remark 6.36. The preceding proof carries through unchanged if A is replaced by an arbitrary matrix, and if the difference set $(S - S) \setminus \{0\}$ is replaced by the sum set $c_1 \cdot S + \cdots + c_l \cdot S$ with $c_1 + \cdots + c_l = 0$ (replace $\boldsymbol{x_j} - \boldsymbol{y_j} \in S - S$ by $c_1\boldsymbol{x_j} + (c_2 + \cdots + c_l)\boldsymbol{y_j} \in c_1 \cdot S + \cdots + c_l \cdot S$). So a weaker version of Corollary 6.11, where the AIR-sumset is replaced by an ordinary sumset, can also be proved by a simple counting argument, without using the slice rank method.

Remark 6.37. Now consider once again the sumset $c_1 \cdot S + \cdots + c_l \cdot S$, but this time assume that $c_1 + \cdots + c_l \neq 0$. In this case, the techniques from this chapter do not say anything non-trivial about the problem of finding a non-trivial k-AP in the sum set $c_1 \cdot S + \cdots + c_l \cdot S$. (But the results from this chapter were later superseded by another paper by Gijswijt [Gij21], and it follows from the results contained therein that Corollary 6.12 remains valid when $c_1 + \cdots + c_l \neq 0$.)

We explain why the results from this chapter do not work when $c_1 + \cdots + c_l \neq 0$. It is tempting to try to repeat the proof of Corollary 6.11, but we run into a problem: The column equivalence classes no longer sum to zero, so we have to replace Theorem 6.9(ii) by Theorem 6.9(i). However, this imposes two extra conditions on the original $m \times k$ matrix in the proof of Corollary 6.11, namely that $A\mathbb{1} = 0$ (i.e. $(\star)$ is balanced) and that $k = \operatorname{rank}(A) + 1$. So we can only say something for a very specific class of linear systems. In fact, this class is so specific that the coefficient matrix must satisfy $\ker(A) = \operatorname{span}(\mathbb{1})$, so every solution of the original system must be constant!

Likewise, it is tempting to try to repeat the proof of Corollary 6.11, but this time replacing Theorem 6.9(ii) by Theorem 6.8(i). After all, to find (say) a non-trivial

[9] To prove this, it is sufficient to note that there exist non-trivial k-APs in $\mathbb{F}_q^n \setminus \{0\}$.

k-AP, it is enough to find a solution with $\boldsymbol{y_1}, \dots, \boldsymbol{y_k}$ pairwise distinct instead of a generic solution. Here we run into another problem. In the proof of Corollary 6.11, we can find a solution $(\boldsymbol{x_1}, \dots, \boldsymbol{x_{lk}}) \in S^{lk}$ of the extended system $(\star')$ with $\boldsymbol{x_1}, \dots, \boldsymbol{x_{lk}}$ pairwise distinct. But when we recombine these to form a solution $(\boldsymbol{y_1}, \dots, \boldsymbol{y_k}) \in (c_1 \cdot S + \dots + c_l \cdot S)^k$ of the original system $(\star)$, we may end up with $\boldsymbol{y_1} = \dots = \boldsymbol{y_k}$, since we have no way to avoid these additional equations. In fact, if we use the proof of Theorem 6.8(i) as an algorithm to find the $\boldsymbol{x_1}, \dots, \boldsymbol{x_{lk}}$, then this is guaranteed to happen: We start with a solution where all variables $\boldsymbol{x_1}, \dots, \boldsymbol{x_{lk}}$ are equal, and then modify the variables in such a way that the contribution to each column equivalence class remains the same, so the equation $\boldsymbol{y_1} = \dots = \boldsymbol{y_k}$ is maintained throughout the proof. Once again, the techniques from this chapter are unable to say anything non-trivial.

The systems studied by Mimura and Tokushige

In a series of papers [MT19a, MT19b, MT20], Mimura and Tokushige studied several specific (classes of) linear systems, and showed that each of them is moderate. These were the first results of this type. We show that our results and techniques furnish alternative proofs for all systems studied by Mimura and Tokushige (though our constants might not be as good).

The systems studied by Mimura and Tokushige have integer entries, and can therefore be interpreted as a linear system over $\mathbb{F}_q$ for an arbitrary prime power $q = p^s$. Depending on the system, Mimura and Tokushige sometimes had to assume that $p \neq 2$ or $p \neq 3$, and we shall do the same.

Example 6.38. In [MT19a], Mimura and Tokushige studied a *star of k three-term arithmetic progressions*, given by the linear system $(\mathcal{S}_{*k})$ with coefficient matrix

$$\begin{pmatrix} 1 & 1 & 0 & 0 & \cdots & 0 & 0 & -2 \\ 0 & 0 & 1 & 1 & \cdots & 0 & 0 & -2 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & 1 & -2 \end{pmatrix} \in \mathbb{F}_q^{k \times (2k+1)},$$

and proved that this system is moderate whenever $p \geq 3$.

This result can be recovered as a special case of Theorem 6.8, and strengthened to $(\mathcal{S}_{*k})$ being temperate by Theorem 6.9. Indeed, $(\mathcal{S}_{*k})$ is a type (RC) linear system, as it is balanced and there is only one column equivalence class of size 1. If $p \neq 2$, then the system is non-degenerate and irreducible, and all column equivalence classes have sum $\pm 2 \neq 0$, so it follows from Theorem 6.8(i) that $(\mathcal{S}_{*k})$ is moderate. Additionally, since there are k equations and $k+1$ column equivalence classes, it follows from Theorem 6.9(i) that $(\mathcal{S}_{*k})$ is temperate. △

Example 6.39. Also in [MT19a], Mimura and Tokushige point out that their proof also extends to a 'fan' of k three-term arithmetic progressions, given by the linear

system $(\mathcal{S}'_{*k})$ with coefficient matrix

$$\begin{pmatrix} 1 & -2 & 0 & 0 & \cdots & 0 & 0 & 1 \\ 0 & 0 & 1 & -2 & \cdots & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 \end{pmatrix} \in \mathbb{F}_q^{k\times(2k+1)}.$$

Analogously to Example 6.38, it follows from Theorem 6.8(i) and Theorem 6.9(i) that $(\mathcal{S}'_{*k})$ is moderate and temperate, provided that $p \neq 2$. △

Example 6.40. In [MT19b], Mimura and Tokushige studied the problem of avoiding a 'W shape', and showed that the linear system $(\mathcal{W})$ with coefficient matrix

$$\begin{pmatrix} 1 & -1 & -1 & 1 & 0 \\ 1 & 0 & -2 & 0 & 1 \end{pmatrix} \in \mathbb{F}_q^{2\times 5}$$

is moderate whenever $p \geq 3$.

This is not a type (RC) linear system, since there are 3 column equivalence classes of size 1, so this result cannot be recovered as a special case of Theorem 6.8 or Theorem 6.9.

Nevertheless, our techniques from §6.3 can be adapted to recover this result as well. Indeed, let Γ_q be the constant from Theorem 6.17, and let $S \subseteq \mathbb{F}_q^n$ with $|S| \geq 4 \cdot (\Gamma_q)^n$. By repeatedly finding a non-trivial 3-AP and removing it from S, we can find a list $\{(\boldsymbol{x_1^{(i)}}, \boldsymbol{x_3^{(i)}}, \boldsymbol{x_5^{(i)}})\}_{i=1}^{L}$ of $L \geq (\Gamma_q)^n$ pairwise disjoint non-trivial 3-APs in S^3. For all $i \in [L]$, set $\boldsymbol{x_2^{(i)}} = \boldsymbol{x_3^{(i)}}$ and $\boldsymbol{x_4^{(i)}} = \boldsymbol{x_5^{(i)}}$, so that $(\boldsymbol{x_1^{(i)}}, \boldsymbol{x_2^{(i)}}, \boldsymbol{x_3^{(i)}}, \boldsymbol{x_4^{(i)}}, \boldsymbol{x_5^{(i)}}) \in S^5$ is a solution of $(\mathcal{W})$. Since 2 and 4 belong to the same column equivalence class, it follows from Corollary 6.21 that there are $i \neq i', i''$ such that the 5-tuple $(\boldsymbol{y_1}, \boldsymbol{y_2}, \boldsymbol{y_3}, \boldsymbol{y_4}, \boldsymbol{y_5}) = (\boldsymbol{x_1^{(i)}}, \boldsymbol{x_2^{(i')}}, \boldsymbol{x_3^{(i)}}, \boldsymbol{x_4^{(i'')}}, \boldsymbol{x_5^{(i)}}) \in S^5$ is also a solution of $(\mathcal{W})$. Then $\boldsymbol{y_1}, \boldsymbol{y_3}, \boldsymbol{y_5}$ are pairwise distinct because they stem from the same non-trivial 3-AP, and $\{\boldsymbol{y_1}, \boldsymbol{y_3}, \boldsymbol{y_5}\} \cap \{\boldsymbol{y_2}, \boldsymbol{y_4}\} = \varnothing$ because they stem from disjoint solutions. Finally, note that $\boldsymbol{y_2} \neq \boldsymbol{y_4}$, for otherwise the first equation of $(\mathcal{W})$ would imply that $\boldsymbol{y_1} = \boldsymbol{y_3}$. This shows that $(\mathcal{W})$ is moderate.

With minor modifications, the preceding argument also shows that $(\mathcal{W})$ is temperate. Indeed, by repeating the argument, but using multiple replacement (Corollary 6.30) instead of single replacement (Corollary 6.21), we can make sure that $\boldsymbol{x_2^{(i')}}$ is not in the line through $\boldsymbol{x_1^{(i)}}$, $\boldsymbol{x_3^{(i)}}$ and $\boldsymbol{x_5^{(i)}}$. Then $\dim(\mathrm{aff}(\boldsymbol{x_1^{(i)}}, \boldsymbol{x_2^{(i')}}, \boldsymbol{x_3^{(i)}}, \boldsymbol{x_4^{(i'')}}, \boldsymbol{x_5^{(i)}})) \geq 2$, so it follows from Corollary 6.26 that this solution is generic. △

Example 6.41. In [MT20], Mimura and Tokushige studied the system (T) with coefficient matrix

$$\begin{pmatrix} 1 & -2 & 1 & 0 & 0 \\ 0 & 0 & -2 & 1 & 1 \end{pmatrix} \in \mathbb{F}_q^{2\times 5},$$

and proved that it is moderate whenever $p \geq 3$.

Once again, this result can be recovered as a special case of Theorem 6.8(i), and strengthened to (T) being temperate by Theorem 6.9(i). △

Example 6.42. In [MT20], Mimura and Tokushige studied the class of linear systems (lS_{k+2}). This class is defined as follows: let $k \geq 1$, and let $a_1, \dots, a_{k+2} \in \mathbb{F}_q$ be non-zero such that $a_1 + \cdots + a_{k+2} = 0$. Then (lS_{k+2}) is given by the coefficient matrix

$$\begin{pmatrix} a_1 & \cdots & a_k & a_{k+1} & a_{k+2} & 0 & 0 & \cdots & 0 & 0 \\ a_1 & \cdots & a_k & 0 & 0 & a_{k+1} & a_{k+2} & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_1 & \cdots & a_k & 0 & 0 & 0 & 0 & \cdots & a_{k+1} & a_{k+2} \end{pmatrix} \in \mathbb{F}_q^{l \times (k+2l)}.$$

In [MT20, Thm. 5], Mimura and Tokushige showed that such a system is always moderate. (This contains the linear system (S_1) from [MT20] as a special case.)

This result can be recovered as a special case of Theorem 6.8, and strengthened to (lS_{k+2}) being temperate by Theorem 6.9. Indeed, (lS_{k+2}) is balanced, and it has one column equivalence class of size $k \geq 1$ and l column equivalence classes of size 2, so it is a type (RC) linear system. Furthermore, the system is non-degenerate and irreducible. Note that, if one column equivalence class sums to zero, then all column equivalence classes must sum to zero, so it follows from either Theorem 6.8(i) or Theorem 6.8(ii) that (lS_{k+2}) is moderate. Furthermore, since the number of equations is l and the number of column equivalence classes is $l + 1$, it follows from either Theorem 6.9(i) or Theorem 6.9(ii) that (lS_{k+2}) is temperate. △

Example 6.43. In [MT20], Mimura and Tokushige studied the class of linear systems $(2T_{k,l})$. This class is defined as follows: let $k \geq 1$ and $l \geq 2$, and let $a_1, \dots, a_{k+l} \in \mathbb{F}_q$ be non-zero such that $a_1 + \cdots + a_{k+l} = 0$. Then $(2T_{k,l})$ is given by the coefficient matrix

$$\begin{pmatrix} a_1 & \cdots & a_k & a_{k+1} & \cdots & a_{k+l} & 0 & \cdots & 0 \\ a_1 & \cdots & a_k & 0 & \cdots & 0 & a_{k+1} & \cdots & a_{k+l} \end{pmatrix} \in \mathbb{F}_q^{2 \times (k+2l)}.$$

In [MT20, Thm. 6], Mimura and Tokushige showed that such a system is always moderate. (This contains the linear system (S_2) from [MT20] as a special case.)

This result can be recovered as a special case of Theorem 6.8, and strengthened to $(2T_{k,l})$ being temperate by Theorem 6.9. The argument is analogous to that of Example 6.42. △

Example 6.44. In [MT20], Mimura and Tokushige studied the linear system (S_3^-) with coefficient matrix

$$\begin{pmatrix} 1 & 1 & 1 & 1 & -4 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & -4 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & -4 \end{pmatrix} \in \mathbb{F}_q^{3 \times 10},$$

and proved that it is moderate whenever $p \neq 2$.[10]

[10]The authors don't make the assumption $p \neq 2$ explicit in their proof. This assumption is necessary because the sum of the second and third row of the coefficient matrix is congruent to $\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \end{pmatrix} \pmod 2$. So for $p = 2$ the system cannot be moderate because it forces two variables to be equal.

This result can be recovered as a special case of Theorem 6.8, provided that $p \neq 2, 3$.[11] The results from this chapter are insufficient to determine whether (S_3^-) is temperate, because there are not enough equations to apply Theorem 6.9(i). △

Example 6.45. Finally, in [MT20, Conjecture 1], Mimura and Tokushige conjectured that the system (S_3) with coefficient matrix

$$\begin{pmatrix} 1 & 1 & 1 & 1 & -4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & -4 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & -4 \end{pmatrix} \in \mathbb{F}_q^{3\times 11}$$

is moderate. This is confirmed by our results. If $p \neq 2$, then it follows from Theorem 6.8(i) and Theorem 6.9(i) that (S_3) is moderate and temperate. If $p = 2$, then some of the columns become zero, so they correspond to free variables. After removing those columns, it follows from Theorem 6.8(ii) and Theorem 6.9(ii) that (S_3) is moderate and temperate. △

In summary: in all examples except Example 6.44, we were able to prove that the system is moderate and temperate, thereby strengthening prior results (and proving a conjecture) of Mimura and Tokushige. In Example 6.44, we gave an alternative proof of the fact that the system is moderate, but we were unable to determine whether the system is also temperate.

In Example 6.40, we could not apply Theorem 6.8. Instead, we needed a proof that was adapted to this particular system, using results from §6.3, to furnish an alternative proof that the system is moderate. In all other examples, the fact that the system is moderate follows immediately from Theorem 6.8.

As a final remark, we point out once again that the results from this chapter were later superseded by another paper by Gijswijt [Gij21]. It follows from the results contained therein that all systems from this section are temperate. This includes Example 6.44, which we were unable to settle in this chapter.

[11] If $p = 2$, then there are three column equivalence classes of size 1, so the system is not of type (RC). Furthermore, if $p \in \{2, 3\}$, then there are column equivalence classes of size 2 that sum to zero, but not all column equivalence classes sum to 0, so neither Theorem 6.8(i) nor Theorem 6.8(ii) applies in this case. If $p \notin \{2, 3\}$, then the system is of type (RC) and none of column equivalence classes sums to zero, so Theorem 6.8(i) applies.

Part III

Tensor products of convex cones

Chapter 7

Outline of Part III

In this chapter, we give an outline of Part III of this dissertation.

Part III is based on the paper [Dob20b], and this chapter is based on the introduction (Chapter 1) of that paper.

7.1 Introduction

Convex cones have applications in almost all branches of mathematics, from algebra and geometry to analysis and optimization. Consequently, convex cones have been studied extensively in their own right, and there is a vast body of work on all kinds of geometrical, analytical, and combinatorial properties of convex cones.

In the study of convex cones, just as in any other area of mathematics, it is important to have good ways of creating new objects from old. One such problem which has attracted a lot of attention is the following. Suppose that we are given convex cones E_+ and F_+ in the vector spaces E and F, respectively. Can we then use the data of E_+ and F_+ to somehow construct a natural cone in the tensor product $E \otimes F$? As it turns out, there are multiple ways to do so [Mer64, PS69], just as there are multiple ways to define a norm on the tensor product of two normed spaces [Rya02].

Among all "reasonable" cones in the tensor product $E \otimes F$, there is a smallest and a largest one, which we denote by $E_+ \otimes^\pi F_+$ and $E_+ \otimes^\varepsilon F_+$, respectively. These have come up many times in the literature, motivated by problems in a variety of different fields. We outline a few of these applications:

- In functional analysis, one is often interested in tensor products of various types of spaces (e.g. Banach spaces, C^*-algebras, operator spaces, etc.). Often the two factors come with natural order structure, in which case it is desirable to find a compatible order structure in the tensor product. This is equivalent to finding a tensor product of the positive cones, and so tensor products of convex cones are closely linked to tensor products of ordered (topological) vector spaces.

- In operator theory, the minimal and maximal tensor product of a positive semidefinite cone with an arbitrary cone C correspond to the smallest and largest operator system with C at its ground level. Question surrounding this minimal and maximal operator system have been studied by several authors; for instance, [PTT11, FNT17, HN21]. Furthermore, these questions turn out to be

closely related to questions about matrix convex sets [PSS18, §7], [Sha21, Thm. 9.11] and free spectrahedra [FNT17], topics which have been studied by authors in geometry, optimization, and quantum information theory.

- In theoretical physics, the theory of "general probabilistic theories" (GPTs) forms a new framework which generalizes both classical and quantum probability [Lam17, Mül21, Plá21, ALPP21]. A GPT derives probability from an arbitrary finite-dimensional Archimedean cone with an order unit. Classical (resp. quantum) probability can then be recovered as a special case by taking a simplex (resp. positive semidefinite) cone.

  Given GPTs E and F, the tensor product $E \otimes F$ corresponds to the composite system (E, F). In this setting, the elements of the smallest cone $E_+ \otimes^\pi F_+$ correspond to the separable states, whereas the elements of $E_+ \otimes^\varepsilon F_+ \setminus E_+ \otimes^\pi F_+$ correspond to the entangled states [Plá21, Def. 5.8]. Thus, understanding tensor products of convex cones is crucial to understanding entanglement in GPTs.

- In polyhedral geometry, the minimal and maximal tensor product of two polyhedral cones are closely related to the tensor product and Hom-polytope of the underlying polytopes. Since Hom and tensor are fundamental constructions in the category of polytopes, their properties have been studied in detail in the literature; see for instance [BCG13].

- In approximation theory, tensor products of convex cones come up naturally in the context of multivariate shape preserving interpolation with cone constraints. For a precise description of this problem and its relation to tensor products of convex cones, see [Mul97].

In each of these settings, the underlying construction is just a tensor product of convex cones. This underpins the importance of a systematic study of tensor products of convex cones, and indeed many papers have already been written about this. However, most of the existing literature only focuses on one of two particular cases: lattice cones and finite-dimensional cones. As a result, the literature is divided into two separate lines of investigation, neither of which addresses the problem in full generality.

The first line of investigation comes from functional analysis. In this setting, the focus has mostly been on Riesz spaces and Banach lattices. Although most classical Banach spaces are lattice-ordered, many other interesting classes of ordered vector spaces are not. For example, the self-adjoint part of a C^*-algebra $\mathcal{A}$ is an ordered vector space with a closed, proper and generating cone $\mathcal{A}_+$, but by Sherman's theorem it is lattice-ordered if and only if $\mathcal{A}$ is commutative. This shows that, in a way, restricting one's attention to lattice-ordered spaces is akin to restricting one's attention to commutative C^*-algebras.

The second line of investigation comes from linear algebra, and encompasses the remaining applications from the preceding list. In this setting, research has dealt exclusively with closed, proper and generating cones in finite-dimensional spaces. This is once again a severe limitation, at least from the perspective of analysis, as finite-dimensional spaces are often of limited use there. Furthermore, even in the finite-dimensional case one occasionally encounters cones which are not closed or not

proper. For example, lexicographical cones in a space of dimension at least 2 are never closed, and quotients/projections of closed, proper cones are not guaranteed to be closed either (see e.g. [Dob20a, Example 6.3]).

There has been very little cross-pollination between these two lines of investigation, and very little has been done beyond these two specific cases. In particular, almost nothing is known about tensor products of infinite-dimensional ordered vector spaces which are not lattice-ordered, or about tensor products of cones which are not closed and/or not proper. This disqualifies many cones from consideration, including even standard cones such as the positive semidefinite cone over an infinite-dimensional Hilbert space.

Furthermore, even in the cases that have been studied, many basic properties have not been noted or proved in the existing literature. For instance, whereas mapping properties play an important role in the similar theory of *normed* tensor products, we are not aware of prior papers which establish the mapping properties of the minimal/maximal tensor product of convex cones. Likewise, only partial results are known about properness of the minimal/maximal tensor product of convex cones, or whether the minimal/maximal tensor product preserves faces of the base cones.

Part III of this dissertation aims to develop a general theory of tensor products of convex cones, without any restrictions on the cones or the ambient spaces. By using ideas from both lines of investigation and borrowing additional techniques from the similar theory of *normed* tensor products, we are able to extend known results to the general setting and prove many completely new results.

In the next section, we give a very brief overview of the existing literature. After that, the remainder of this chapter gives a comprehensive overview of the main results of Part III.

7.2 Brief literature overview

The study of tensor products of ordered topological vector spaces was initiated in the 1960s by Merklen [Mer64],[1] Hulanicki and Phelps [HP68], Popa [Pop68, Pop69], and Peressini and Sherbert [PS69]. From the 1970s onwards, the focus has mostly been on Riesz spaces ([Sch72, Fre72, Fre74, Wit74, Sch74, Bir76, FT79, Nie82, GL88, Nie88, GL89, Bla16, ABJ18, BT22, BGY22]) and, in a separate line of investigation, on closed cones in finite-dimensional spaces ([BL75, Bar76, HFP76, Bar78a, Bar78b, Bar81, BLP87, ST90, Tam92, Tam95, Mul97, Hil08, HN21, ALPP21]). For general ordered vector spaces, some of the basic questions remain unanswered (and, on one occasion, escaped from collective memory, as we point out below).

The most comprehensive paper on tensor products of general ordered vector spaces is the article of Peressini and Sherbert [PS69]. It contains an in-depth study of the

[1] It appears that Merklen was the first to study tensor products of ordered vector spaces, but his article is very hard to find, and contains several errors. For instance, [Mer64, Teorema 5] states that the weak closure of the projective cone $E_+ \otimes^\pi F_+$ is a proper cone if at least one of E_+ and F_+ is proper, provided that E_+ and F_+ are weakly closed. Likewise, [Mer64, Teorema 9] states the same for the injective cone. Both of these statements are incorrect, as can be seen by taking $E_+ = \mathbb{R}_{\geq 0}$ and $F_+ = \mathbb{R}$. The correct statement is that both E_+ and F_+ should be proper; see Theorem D and Theorem B below.

properties of the *projective* (minimal) and *injective* (maximal) cone in the tensor product.[2] It answers various topological and order-theoretic questions about these cones, for instance relating to normality and order units. Furthermore, it establishes a few sufficient conditions for the projective/injective cone to be proper, but it does not provide precise necessary and sufficient conditions.

Conditions for the projective cone to be proper were quickly provided by Dermenjian and Saint-Raymond [DS70], but their result seems to have been unknown to later generations of mathematicians. Only recently was this question answered (again) by Wortel [Wor19]; until then only special cases were assumed to be known in the literature. For the injective cone, no precise necessary and sufficient conditions for properness are known in the literature.

The situation is much better in the setting of lattice-ordered or finite-dimensional spaces. For lattice-ordered spaces, a lot has been said about the problem of turning the tensor product (or its completion) into a lattice-ordered space as well [Sch72, Fre72, Fre74, Wit74], and connections between such lattice tensor products and lattices of operators are well-known [Sch74, §IV.7]. However, such results are rather specific to lattice-ordered spaces, and have little hope of being generalized to general (non-lattice-ordered) spaces.

Likewise, in the finite-dimensional setting, much more is known. Here research has focused on cones that are closed, proper and generating. This is sufficient to guarantee that the projective and injective tensor product are closed, proper and generating as well [Tam77b], so finding criteria for properness is not an issue here. More advanced results have been obtained as well; see for instance [BL75, Bar76, Bar81, Tam92]. In particular, in the context of cones of positive operators, Tam gave a construction which can be used to obtain faces in the injective cone from faces of the base cones [Tam92, §4]. We will extend this result; see §7.6.

There is very little overlap between the lattice-ordered and the finite-dimensional theory, because they deal with very different questions. After all, the only finite-dimensional closed lattice cones are the simplex cones (i.e. the ones isomorphic to $\mathbb{R}^n_{\geq 0}$), which are not very interesting from either perspective. However, one problem that has been studied in both settings is the question whether or not the projective cone is dense in the injective cone. Birnbaum [Bir76] showed that this is true whenever E and F are locally convex lattices and gave an example which shows that it is not true in general. Very recently, this problem was settled in the finite-dimensional case by Aubrun, Lami, Palazuelos and Plávala [ALPP21]. They proved that, for closed, proper and generating cones E_+ and F_+ in finite-dimensional spaces E and F, one has $E_+ \otimes^\pi F_+ = E_+ \otimes^\varepsilon F_+$ if and only if at least one of E_+ and F_+ is a simplex cone. Around the same time, we independently found a different proof of this result for nearly all cones [Dob20b], which we have included in this manuscript (see §7.8).

[2] A note on terminology: several authors refer to the maximal cone as the *biprojective cone*. We aim to show that it is in many ways analogous to the injective norm, and as such deserves the name *injective cone*. This term has also occasionally been used before, for instance by Wittstock [Wit74] and Mulansky [Mul97].

7.3 Scope and notation

We now outline the scope of Part III of this dissertation, and we cover the basic notation needed to state our main results in the upcoming sections.

In Part III of this dissertation, we study the projective and injective tensor product of two convex cones $E_+ \subseteq E$ and $F_+ \subseteq F$, where E and F are either real vector spaces or real *topological* vector spaces.

Topological considerations will not matter too much for our investigation, but to build a satisfactory duality theory we need to at least keep track of the duals of all spaces involved. Hence, instead of remembering the topology of a vector space E (or the fact that E has no topology), we only remember the dual pair $\langle E, E' \rangle$ to which E belongs. The advantage of this approach is twofold: it allows us to treat the topological and non-topological cases simultaneously (if E has no topology, let $E' := E^*$ be the algebraic dual), and it allows us to completely ignore any topological issues in the tensor product, thereby sidestepping the notoriously difficult theory of topological tensor products. One downside of this approach is the following: since we have no topology on E, we must occasionally refer to the weak closure $\overline{E_+}^{w}$ of E_+, instead of the ordinary closure. However, we remind the reader that in every locally convex space, the weak closure of a convex set coincides with its original closure.

We now recall some basic notation. A *convex cone* (otherwise known as a *wedge*) in a real vector space E is a non-empty subset $\mathcal{K} \subseteq E$ satisfying $\mathcal{K} + \mathcal{K} \subseteq \mathcal{K}$ and $\lambda \mathcal{K} \subseteq \mathcal{K}$ for all $\lambda \in \mathbb{R}_{\geq 0}$. The *lineality space* of a convex cone $\mathcal{K}$ is the linear subspace $\operatorname{lin}(\mathcal{K}) := \mathcal{K} \cap -\mathcal{K}$. We say that a convex cone $\mathcal{K}$ is *proper* if $\operatorname{lin}(\mathcal{K}) = \{0\}$ and *semisimple* if its weak closure $\overline{\mathcal{K}}^{w}$ is proper.

Let E and F be vector spaces, and let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones. The *projective cone* in $E \otimes F$ is given by

$$E_+ \otimes^{\pi} F_+ := \left\{ \sum_{i=1}^{k} x_i \otimes y_i \ : \ k \in \mathbb{N},\ x_1, \ldots, x_k \in E_+,\ y_1, \ldots, y_k \in F_+ \right\}.$$

Furthermore, if E and F belong to the dual pairs $\langle E, E' \rangle$ and $\langle F, F' \rangle$, then the *injective cone* in $E \otimes F$ is given by

$$E_+ \otimes^{\varepsilon} F_+ := \big\{ u \in E \otimes F \ : \ \langle u, \varphi \otimes \psi \rangle \geq 0 \text{ for all } \varphi \in E'_+,\ \psi \in F'_+ \big\}.$$

For additional notation, see Chapter 8, or refer to the glossary of notation on page 201.

A note about cones in the completed tensor product

So far, the study of tensor products of convex cones has mostly been limited to cones in the algebraic tensor product, with the exception of some results on tensor products of Banach lattices. However, the algebraic tensor product is often of limited use in analysis; instead, one is usually interested in its completion with respect to some suitable topology. For this reason, we also aim to initiate a study of the projective and injective cones in completed locally convex tensor products.

When dealing with *topological* tensor products, one has to define the topology *before* taking the completion, for obviously the completion depends on the chosen topology.

On the other hand, the cone is unrelated to the topology, and can therefore be defined directly on the completion. This gives rise to a natural extension of the injective cone to the completed tensor product, which we will also study in this dissertation. On the other hand, the projective cone in the completed tensor product $E \tilde{\otimes}_\alpha F$ is merely the same cone embedded in a larger ambient space[3], so there is little reason to study this cone separately.

An overview of the cones under consideration, their notation, and their domains of definition, is given in Table 7.1. (In all cases, $E_+ \subseteq E$ and $F_+ \subseteq F$ are convex cones in the primal spaces.)

Table 7.1: The domain of definition of the projective/injective cones studied in Part III of this dissertation.

Cone	Ambient space	Notation	Domain of definition
Projective	$E \otimes F$	$E_+ \otimes^\pi F_+$	E and F vector spaces
Injective	$E \otimes F$	$E_+ \otimes^\varepsilon F_+$	$\langle E, E' \rangle$ and $\langle F, F' \rangle$ dual pairs
Injective	$E \tilde{\otimes}_\alpha F$	$E_+ \tilde{\otimes}^\varepsilon_\alpha F_+$	E and F complete lcs; α a compatible lc topology on $E \otimes F$

In the remainder of this chapter, we state the main results of Part III only for cones in the algebraic tensor product. Similar results hold for the injective cone in completed locally convex tensor products, but these are harder to state, as they often require additional (topological) assumptions. Precise statements can be found in Chapter 10 on the injective cone.

7.4 Mapping properties

In the theory of normed tensor products, it is well-known that the projective norm preserves metric surjections (quotients) and the injective norm preserves metric injections (isometries), and these simple mapping properties play an important role in the theory. By looking at the corresponding types of positive linear maps, we show that the projective and injective cones have analogous mapping properties.

Let E and F be vector spaces, and let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones. We say that a linear map $T \in \mathrm{L}(E, F)$ is *positive* if $T[E_+] \subseteq F_+$, a *pullback* (or *bipositive operator*) if $E_+ = T^{-1}[F_+]$, and a *pushforward* if $T[E_+] = F_+$. Furthermore, if E and F belong to dual pairs $\langle E, E' \rangle$ and $\langle F, F' \rangle$, then we say that an operator $T \in \mathfrak{L}(E_w, F_w)$ is an *approximate pullback* (or *approximately bipositive*) if $\overline{E_+}^w = T^{-1}[\overline{F_+}^w]$, and an *approximate pushforward* if $\overline{T[E_+]}^w = \overline{F_+}^w$. (Recall that in a locally convex space, the weak closure of a convex set coincides with its original closure.) Every pushforward is also an approximate pushforward, but a pullback is not necessarily an approximate pullback (see §8.3).

[3]We define the projective cone algebraically, without taking its closure. This is the prevalent definition in the literature, but might not be appropriate for all applications. We do prove a few results about its closure; see Corollary I and Theorem D.

A typical example of a pullback is an order embedding: if (E, E_+) is order isomorphic to a subspace of (F, F_+), then the embedding $E \hookrightarrow F$ is a pullback. A typical example of a pushforward is a quotient or projection.

In the normed theory, the projective norm preserves metric surjections (quotients) and the injective norm preserves metric injections (isometries). We prove a similar result for cones:

Theorem A. *The projective cone preserves positive linear maps, (approximate) pushforwards, and order retracts, but not (approximate) pullbacks.*

The injective cone preserves weakly continuous positive linear maps, approximate pullbacks, and topological order retracts, but not pullbacks or (approximate) pushforwards.

In particular, the injective cone preserves order embeddings when the cones are weakly closed. We believe this result to be new, even in the finite-dimensional setting.

The proof of Theorem A will be given in §9.2 (projective cone) and §10.2 (injective cone). An overview of these mapping properties is given in Table 7.2.

Table 7.2: Types of maps preserved by the projective/injective cone.

Type of map	Preserved by	
	Projective cone	Injective cone
Positive map	✓	✓
Pushforward	✓	
Approximate pushforward	✓	
Pullback		
Approximate pullback		✓
Retract (positive projection)	✓	✓

Note that the injective cone only preserves *approximate* pullbacks. It is not so strange that it does not preserve all pullbacks: the injective cone does not see the difference between E_+ and $\overline{E_+}^w$, and a pullback for E_+ is not necessarily a pullback for $\overline{E_+}^w$ (for details, see §10.2). In general, the properties of the injective cone depend on those of $\overline{E_+}^w$ and $\overline{F_+}^w$ rather than E_+ and F_+. By contrast, the *projective* cone does see the difference between E_+ and $\overline{E_+}^w$, so it preserves both pushforwards and approximate pushforwards.

7.5 Criteria for properness, the lineality space, and semisimplicity

Another basic question about tensor products of convex cones is to determine when the projective or injective cone is proper. Peressini and Sherbert [PS69] found a few sufficient conditions, but their paper does not specify precise necessary and sufficient

criteria. For the projective cone, precise conditions were found by Dermenjian and Saint-Raymond [DS70], and rediscovered in recent years by Wortel [Wor19].[4] For the injective cone, no such result is known, except in the finite-dimensional case. In this dissertation, we give a simpler proof for the projective cone, and we also settle the problem for the injective cone.

Theorem B. *The projective cone $E_+ \otimes^\pi F_+$ is proper if and only if $E_+ = \{0\}$, or $F_+ = \{0\}$, or both E_+ and F_+ are proper cones (cf. [DS70, Théorème 2]).*
The injective cone $E_+ \otimes^\varepsilon F_+$ is proper if and only if $E = \{0\}$, or $F = \{0\}$, or both $\overline{E_+}^w$ and $\overline{F_+}^w$ are proper cones.

The proof of Theorem B will be given in §9.3 (projective cone) and §10.3 (injective cone). Note that there is a subtle difference between the corner cases in Theorem B: the corner case for the projective cone is when one of the *cones* is trivial, whereas the corner case for the injective cone is when one of the *spaces* is trivial. In partial explanation of this discrepancy, we establish direct formulas for the lineality spaces, from which the criteria of Theorem B can easily be recovered.

Theorem C. *The lineality space of the projective/injective cone is*

$$\lin(E_+ \otimes^\pi F_+) = (\lin(E_+) \otimes \operatorname{span}(F_+)) + (\operatorname{span}(E_+) \otimes \lin(F_+));$$
$$\lin(E_+ \otimes^\varepsilon F_+) = (\lin(\overline{E_+}^w) \otimes F) + (E \otimes \lin(\overline{F_+}^w)).$$

The proof of Theorem C will be given in Corollary 9.17 (projective cone) and Corollary 10.37 (injective cone).

We also address the related question of finding precise necessary and sufficient conditions for the *closure* of the projective cone to be proper. (For the injective cone, this is already addressed by Theorem B, because the injective cone is always closed.) Recall that we say that E_+ is *semisimple* if its weak closure is proper. If E is locally convex, then this is the same as saying that its ordinary closure is proper, because the weak and original closure of a convex set in a locally convex space coincide. We prove the following semisimplicity version of Theorem B.

Theorem D. *The projective cone $E_+ \otimes^\pi F_+$ is semisimple if and only if $E_+ = \{0\}$, or $F_+ = \{0\}$, or both E_+ and F_+ are semisimple.*

A parallel result was proved by van Gaans and Kalauch [GK10]: if E_+ and F_+ are Archimedean, then their projective tensor product is contained in an Archimedean proper cone. Neither of these two results implies the other.

The proof of Theorem D will be given in §11.3.

We also address the question of semisimplicity in completed locally convex tensor products. In §11.4, we prove that the injective cone remains semisimple in the completed injective tensor product, and more generally, in every completion $E \mathbin{\tilde{\otimes}_\alpha} F$ for which the natural map $E \mathbin{\tilde{\otimes}_\alpha} F \to E \mathbin{\tilde{\otimes}_\varepsilon} F$ is injective. However, we do not know whether the projective cone remains semisimple in the completed projective tensor product

[4] The result of Dermenjian and Saint-Raymond seems to have been unknown to later generations of mathematicians, and until recently only special cases were assumed to be known in the literature.

$E \tilde{\otimes}_\pi F$; see Question 11.16. This question is related to the approximation property, as we will explain in §11.4.

7.6 Faces and extremal rays

Next, we turn our attention to another permanence property: preservation of faces. In the normed theory, it follows from a result of Tseitlin [Tse76] that the projective norm sometimes preserves extreme points of the closed unit ball, provided that certain topological requirements are met. Further results in this direction are known for stronger notions of extreme points, such as denting points [RS86b, Wer87]. This leads us to ask to which extent the projective and injective cones preserve extremal rays, or more generally, faces.

In the infinite-dimensional setting, we are not aware of prior results in this direction. For closed cones in finite-dimensional spaces, some constructions are known in the literature. The injective cone $E_+ \otimes^\varepsilon F_+$ can be interpreted as a cone of positive operators $E^* \to F$, whose faces have already been studied by many authors. However, some information is lost in passing from E to E^*, so instead we give a different construction which we believe to be more natural, and we extend this to the general setting. For the projective cone, Tam [Tam92, § 4] pointed out one way to construct faces (without proof). We extend this construction to the general setting, give a full proof, and show that a pair of faces of the base cones give rise to not one but four natural faces of the projective tensor product.

Faces of the projective cone

By combining the mapping properties with the properness criteria, we can show that the projective cone preserves faces.

Theorem E. *If $M \subseteq E_+$ and $N \subseteq F_+$ are faces, then $(M \otimes^\pi F_+) + (E_+ \otimes^\pi N)$ and $(M \otimes^\pi N) + \operatorname{lin}(E_+ \otimes^\pi F_+)$ are faces of the projective cone $E_+ \otimes^\pi F_+$.*

In particular, if E_+ and F_+ are proper cones, then $M \otimes^\pi N$ is a face of $E_+ \otimes^\pi F_+$.

For closed, proper and generating cones in finite-dimensional spaces, this last property was already noted (without proof) by Tam in [Tam77a, p. 53] and [Tam92, p. 71]. We have recovered his simple proof for this special case (see Remark 9.20), but a different technique is needed to prove the general case. In the full generality stated here, Theorem E is a non-trivial result which contains Theorem B as a special case (by setting $M = N = \{0\}$). Remarkably, it is true without any niceness assumptions on the cones E_+ and F_+ or the faces M and N.

The proof of Theorem E will be given in §9.4. There we also mention two other faces induced by M and N, showing that a pair of faces $M \subseteq E_+$ and $N \subseteq F_+$ gives rise to not one but four natural faces of the projective cone. This is a new result, even in the finite-dimensional case.

As an application of Theorem E, we prove that the tensor product of symmetric convex sets preserves proper faces.

Theorem F. *Let E and F be real vector spaces, let $C \subseteq E$, $D \subseteq F$ be absolutely convex, and let $M \subset C$, $N \subset D$ be proper faces. Then* $\operatorname{conv}(M \otimes_s N)$ *is a face of* $\operatorname{conv}(C \otimes_s D)$.

For extreme points, results in this direction were known in the setting of normed tensor products (where C and D are the closed unit balls of the norms of E and F). However, in that setting, stronger assumptions are needed (E and F must be Banach spaces such that at least one of E and F has the approximation property and at least one of E and F has the Radon–Nikodym property), and a stronger conclusion is obtained ($x \otimes y$ is an extreme point of the *closure* of $\operatorname{conv}(C \otimes_s D)$ in the *completed* projective Banach space tensor product); see our remarks following Corollary 9.31. To our knowledge, no such results are known for higher faces, and we are not aware of a general statement like Theorem F in the literature.

Note that Theorem F is a purely algebraic statement, as we do not take closures. We do not know whether it remains true after taking closures, but we suspect it does not (see Remark 9.32). However, if C and D are compact, then $\operatorname{conv}(C \otimes_s D)$ is compact as well, so in particular it follows from Theorem F that the projective norm preserves proper faces of the closed unit ball of finite-dimensional spaces (Corollary 9.31). As far as we know, this had only been known for extreme points.

The proof of Theorem F will be given in §9.6.

Ideals for the injective cone

The injective cone $E_+ \otimes^\varepsilon F_+$ can be interpreted as a subcone of the cone of positive operators $E' \to F$. Since there has been a lot of research into the properties of such cones, a lot has already been said about their faces. For every set $M' \subseteq E'_+$ and every face $N \subseteq F_+$, it is trivially easy to show that the set of positive operators $T \in \mathfrak{L}(E', F)$ satisfying $T[M'] \subseteq N$ forms a face (see Lemma 10.27), so this gives us a plethora of faces in the injective cone. On the other hand, finding all extremal rays of the cone of positive operators is a notoriously difficult problem, so the face structure of the injective cone is still far from fully understood.

Although it is not so hard to construct faces of the injective cone from faces of the base cones, it is unclear what the "right" way of doing so is. Only interpreting $E_+ \otimes^\varepsilon F_+$ as a cone of positive operators $E' \to F$ is a bit unsatisfactory; we might just as well have interpreted it as a cone of positive operators $F' \to E$. Apart from the fact that this is not a symmetric formulation, this poses a bigger problem: in both interpretations, we have one primal and one dual space, but faces are not well-behaved under duality (not every face of E'_+ is the dual of a face of E_+ and vice versa). As far as we know, this problem has not been addressed in the literature, where the focus has been on cones of positive operators $E \to F$ instead of injective tensor products.

In this dissertation, we set out to prove a satisfactory injective counterpart of Theorem E. To do so, we believe we should change perspective from faces to ideals. An (*order*) *ideal* in a preordered vector space (E, E_+) is a subspace $I \subseteq E$ for which the quotient cone $(E/I)_+$ is proper (for other equivalent definitions, see Proposition A.2). There is a close relationship between faces and ideals: the map $I \mapsto I_+$ $(= I \cap E_+)$

defines a surjective many-to-one correspondence between the order ideals of the preordered vector space (E, E_+) and the faces of E_+ (see Appendix A.1).

The benefit of working with ideals instead of faces is twofold. First, in the infinite-dimensional (topological) setting, it is often important to work with *closed* ideals so as to have a useable quotient, but it is not always easy (or even possible) to tell from a face whether or not it occurs as the positive part of a closed ideal. Second, whereas faces of the injective cone can only be given by implicit formulas (all positive operators mapping certain sets into certain faces), for ideals we get the following very simple explicit formulas.

Theorem G. *If $I \subseteq E$ and $J \subseteq F$ are ideals with respect to $\overline{E_+}^w$ and $\overline{F_+}^w$, then $(I \otimes J) + \operatorname{lin}(E_+ \otimes^\varepsilon F_+)$ is an ideal with respect to the injective cone $E_+ \otimes^\varepsilon F_+$.*

Additionally, if I is weakly closed and $(E/I)_+$ is semisimple, or if J is weakly closed and $(F/J)_+$ is semisimple, then $(I \otimes F) + (E \otimes J)$ is also an ideal with respect to the injective cone.

We believe Theorem G to be new, even in the finite-dimensional case. Note that the first formula simplifies to $I \otimes J$ whenever $\overline{E_+}^w$ and $\overline{F_+}^w$ are proper (by Theorem B or Theorem C).

The proof of Theorem G will be given in §10.5. There we will also show that the extra assumption that $(E/I)_+$ or $(F/J)_+$ is semisimple cannot be omitted from the second part of Theorem G (see Example 10.35). Furthermore, we will extend Theorem G to completed locally convex tensor products, though this requires additional topological assumptions (see Theorem 10.46 and Theorem 10.47).

Contrary to the projective case, the preceding results do not have an application to tensor products of symmetric convex sets. The injective analogue of Theorem F is simply not true, because the injective norm does not preserve extreme points of the unit balls (see Remark 10.52). This makes it all the more remarkable that the injective cone preserves faces and extremal rays.

Extremal rays

As a special case of Theorem E and Theorem G, we show that the projective and injective cones preserve extremal rays.

Theorem H. *A vector $u \in E \otimes F$ is an extremal direction of the projective cone $E_+ \otimes^\pi F_+$ if and only if u can be written as $u = x \otimes y$, where x and y are extremal directions of E_+ and F_+.*

If x and y are extremal directions of $\overline{E_+}^w$ and $\overline{F_+}^w$, then $x \otimes y$ is an extremal direction of the injective cone $E_+ \otimes^\varepsilon F_+$. All extremal directions of (tensor) rank one are of this form, but there may also be extremal directions of larger rank.

For closed, proper and generating cones in finite-dimensional spaces, this was already known; see for instance [HFP76, Thm. 3.4(2)][5] or [Tam95, Thm. 3.1]. We

[5] Their result is formulated only for polyhedral cones, but the proof also works for other closed, proper and generating cones.

extend it to arbitrary cones, and also to cones in completed locally convex tensor products (Corollary 10.50).

The proof of Theorem H will be given in §9.5 (projective cone) and §10.6/§11.1 (injective cone). An immediate consequence is that every "reasonable crosscone" (see Chapter 11) preserves extremal rays whenever E_+ and F_+ are weakly closed.

Corollary I. *If E_+ and F_+ are weakly closed, and if x and y are extremal directions of E_+ and F_+, then $x \otimes y$ is an extremal direction of every convex cone $\mathcal{K} \subseteq E \otimes F$ with $E_+ \otimes^\pi F_+ \subseteq \mathcal{K} \subseteq E_+ \otimes^\varepsilon F_+$.*

In particular, the closure of the projective cone also preserves extremal rays. We do not know if it also preserves higher faces (as in Theorem E). For further discussion of this problem, see Chapter 14. For more on reasonable crosscones, see Chapter 11.

7.7 Special properties in the finite-dimensional case

In the linear algebra literature, all papers on tensor products of convex cones have focused on closed, proper and generating cones in finite-dimensional spaces. On the other hand, in the functional analysis literature, most of the focus has been on tensor products of Archimedean lattice cones. Therefore the only cones which are covered by both regimes are the ones isomorphic to the standard cone $\mathbb{R}^n_{\geq 0}$ (all Archimedean lattice cones in $\mathbb{R}^n$ are isomorphic to $\mathbb{R}^n_{\geq 0}$), which are not very interesting from either perspective. Consequently, these two lines of investigation have focused on completely different problems.

Even if one is primarily interested in tensor products of infinite-dimensional convex cones, it is good to be aware of the finite-dimensional theory, as various fundamental phenomena can already be observed here. For this reason, in Chapter 12, we give an overview of the most important additional properties in the finite-dimensional setting (with closed cones).

The main results of Chapter 12 are threefold. First, for closed cones E_+ and F_+ in finite-dimensional spaces, we show that the projective cone $E_+ \otimes^\pi F_+$ can be interpreted as the cone of positive operators $E^* \to F$ that factor positively through some finite-dimensional Archimedean Riesz space (i.e. though some $\mathbb{R}^n$ with the standard cone $\mathbb{R}^n_{\geq 0}$). Second, we show that the closure of the projective cone $E_+ \otimes^\pi F_+$ is equal to the projective cone $\overline{E_+} \otimes^\pi \overline{F_+}$, thereby extending a result of Tam [Tam77b], who proved this in the case that E_+ and F_+ are closed, proper and generating. Third, we study the basic properties of order retracts of finite-dimensional cones, and give many examples of retracts occurring in standard cones.

7.8 Many examples where the projective and injective cone differ

Another question which has attracted a lot of attention is to determine under which circumstances the projective cone $E_+ \otimes^\pi F_+$ is dense in the injective cone $E_+ \otimes^\varepsilon F_+$. For locally convex lattices E and F, Birnbaum [Bir76, Prop. 3] proved that $E_+ \otimes^\pi F_+$

is dense in $E_+ \otimes^\varepsilon F_+$ in the projective topology (and therefore in every coarser topology), and followed this by an example showing that this is not true for all ordered locally convex spaces (not necessarily lattice ordered). In general, however, the infinite-dimensional version of this problem does not appear to be well understood.

A lot more is known in the finite-dimensional setting (with closed, proper and generating cones), where various results in this direction have been obtained since the 1970s. Here $E_+ \otimes^\pi F_+$ is automatically closed whenever E_+ and F_+ are closed (by the results from §7.7), so the question is whether or not the projective and injective cones are equal.

Let E and F be finite-dimensional spaces, and let $E_+ \subseteq E$, $F_+ \subseteq F$ be closed, proper and generating convex cones. We say that E_+ is a *simplex cone* (or *Yudin cone*) if it is generated by a basis (or equivalently, if there is a linear isomorphism $E \cong \mathbb{R}^n$ that identifies E_+ with the standard cone $\mathbb{R}^n_{\geq 0}$).

In the 1970s, Barker showed that $E_+ \otimes^\pi F_+ = E_+ \otimes^\varepsilon F_+$ whenever E_+ or F_+ is a simplex cone [Bar76], and conversely conjectured that E_+ or F_+ must be a simplex cone whenever $E_+ \otimes^\pi F_+ = E_+ \otimes^\varepsilon F_+$ [Bar81, p. 277]. His conjecture remained open for a very long time, but partial results were obtained by Barker and Loewy [BL75, Prop. 3.1], who proved the conjecture when $F_+ = E_+^*$, and by Poole [Poo75, Thm. 5.15], who proved it when E_+ and F_+ are polyhedral. More recently, Huber and Netzer [HN21] proved the conjecture when E_+ is a positive semidefinite cone and F_+ is a polyhedral cone (or vice versa).

In Chapter 13, we prove Barker's conjecture for nearly all[6] pairs (E_+, F_+) of closed, proper and generating cones in finite-dimensional spaces. Recall that a closed, proper and generating convex cone is called *strictly convex* if every non-zero boundary point is an extremal direction, and *smooth* if every non-zero boundary point has exactly one supporting hyperplane. It is well-known that E_+ is strictly convex if and only if E_+^* is smooth, and vice versa. We prove Barker's conjecture in the case that $\dim(E) \geq \dim(F)$ and E_+ is smooth or strictly convex.

Theorem J. *Let E, F be finite-dimensional real vector spaces, and let $E_+ \subseteq E$, $F_+ \subseteq F$ be closed, proper, and generating convex cones. If $\dim(E) \geq \dim(F)$, and if E_+ is strictly convex or smooth, then one has $E_+ \otimes^\pi F_+ = E_+ \otimes^\varepsilon F_+$ if and only if F_+ is a simplex cone.*

The set of convex bodies in $\mathbb{R}^n$ which are not smooth or strictly convex is meagre in the Hausdorff metric [Kle59], and even satisfies the stronger notion of "σ-porosity" [Zam87]. As such, Theorem J shows that the projective and injective cone differ for nearly all[6] pairs of closed, proper, and generating cones (E_+, F_+).

The proof of Theorem J will be given in §13.3.

Although Theorem J covers nearly all cones, it does not cover most standard cones. For instance, polyhedral cones and positive semidefinite cones (and their duals) have many non-trivial faces, so they are not smooth or strictly convex. We complement Theorem J with a similar result for combinations of standard cones.

[6]The term 'nearly all' has a precise meaning (namely, up to a σ-porous set); see for instance [Zam87].

Theorem K. *Let E, F be finite-dimensional real vector spaces, and let $E_+ \subseteq E$, $F_+ \subseteq F$ be closed, proper, and generating convex cones. Assume that each of E_+ and F_+ is one of the following (all combinations allowed):*

(i) a polyhedral cone;

(ii) a second-order cone;

(iii) a (real or complex) positive semidefinite cone.

Then one has $E_+ \otimes^\pi F_+ = E_+ \otimes^\varepsilon F_+$ if and only if at least one of E_+ and F_+ is a simplex cone.

Using retracts, Theorem K could already be deduced from the aforementioned known results of Poole [Poo75, Thm. 5.15] and Huber and Netzer [HN21]. We give a new proof of Theorem K, thereby also providing new proofs of the results of Poole and of Huber and Netzer.

The proof of Theorem K will be given in §13.4.

Remark L. As the manuscript [Dob20b] that forms the basis for Part III of this dissertation was being written, the preceding results were superseded by independent work of Aubrun, Lami, Palazuelos and Plávala [ALPP21]. Motivated by questions in theoretical physics, they proved that $E_+ \otimes^\pi F_+ = E_+ \otimes^\varepsilon F_+$ if and only if at least one of E_+ and F_+ is a simplex cone (provided that E_+ and F_+ are closed, proper, and generating). Both Theorem J and Theorem K are special cases of this result.

The proofs in Part III of this dissertation were discovered independently around the same time, and our proofs differ significantly from the proof in [ALPP21]. Although we recover their result for nearly all cones, we have not been able to recover it in full generality.

Applications to operator systems

The recent resurgence of interest in the question of whether or not $E_+ \otimes^\pi F_+ = E_+ \otimes^\varepsilon F_+$ is due in part to recent developments in the study of operator systems [FNT17, HN21]. Reformulated in terms of operator systems (using notation from [FNT17]), our results prove the following.

Corollary M. *Let $C \subseteq \mathbb{R}^d$ be a closed, proper, and generating convex cone. If $d \leq 4$, or if C is strictly convex, or smooth, or polyhedral, or (real or complex) positive semidefinite, then the following are equivalent:*

(i) C is a simplex cone;

(ii) the minimal and maximal operator systems $C^{\min}$ and $C^{\max}$ are equal;

(iii) there exists $n \geq 2$ for which $C_n^{\min} = C_n^{\max}$;

(iv) one has $C_2^{\min} = C_2^{\max}$.

Again, Corollary M was superseded by the work of Aubrun, Lami, Palazuelos and Plávala [ALPP21], who removed the additional assumptions on the cone C (see Remark L).

The proof of Corollary M will be given in §13.4.

7.9 Appendix: faces and ideals

The main body of Part III is complemented by an appendix on faces and ideals of convex cones in infinite-dimensional spaces. This material is not directly related to tensor products, but will be used extensively in the proofs.

Although faces and ideals have each received a lot of attention in the literature, the link between these concepts does not appear to be well-known. The relationship is very simple: the map $I \mapsto I_+$ defines a surjective many-to-one correspondence between ideals and faces (Appendix A.1). Going back and forth between faces and ideals is crucial in our study of the faces of the projective/injective cone.

In Appendix A, we study the basic properties of faces and ideals and the connection between the two (Appendix A.1), we outline to which extent the homomorphism and isomorphism theorems hold (Appendix A.2), and we study dual and exposed faces in infinite-dimensional cones (Appendix A.3). A more detailed outline will be given at the beginning of Appendix A.

7.10 Organization of Part III

In Chapter 8, we recall all required notation and terminology for Part III. This is complemented by a glossary notation and an index, both of which can be found at the end of this dissertation.

In Chapter 9, we study the properties of the projective cone. Here we prove all of the main results for the projective cone (see §7.4–7.6), with the exception of Theorem D, whose proof is deferred until Chapter 11.

Likewise, in Chapter 10, we study the properties of the injective cone. Here we prove all of the main results for the injective cone (see §7.4–7.6).

In Chapter 11, we study the basic properties of the so-called 'reasonable crosscones' (that is, arbitrary cones which lie somewhere between the projective and injective cone). We show that all reasonable crosscones have the same rank 1 tensors whenever E_+ and F_+ are weakly closed and proper, and we briefly look at ideals and extremal rays of reasonable crosscones. Furthermore, we study semisimplicity of reasonable crosscones, and we give a proof of Theorem D (on the semisimplicity of the projective cone). We also look at questions surrounding semisimplicity of cones in *completed* locally convex tensor products, and we discuss how these questions are related to topological issues and the approximation property.

In Chapter 12, we give an overview of the most important additional properties in the finite-dimensional setting (see §7.7). Building on this, in Chapter 13, we give many (finite-dimensional) examples where the projective and injective cone are different (see §7.8).

In Chapter 14, we discuss a few open problems related to Part III.

Finally, in Appendix A, we discuss the relation between faces and order ideals. These results are not very well known, and will be used extensively in the main body of Part III, so we have included them for completeness.

Chapter 8

Preliminaries for Part III

This chapter covers the prerequisites for Part III, including: topological vector spaces, dual pairs, convex cones, and ordered vector spaces.

This chapter is based on Chapter 2 of [Dob20b].

Introduction

In Part III, we study tensor products of convex cones E_+, F_+ in real vector spaces E, F. Occasionally, E and F will be *topological* vector spaces, but usually they are only assumed to be the primal spaces of the dual pairs $\langle E, E' \rangle$, $\langle F, F' \rangle$. In this chapter, we cover the necessary terminology and notation for Part III. This is complemented by a glossary of notation and an index, both of which can be found at the end of this dissertation.

8.1 Topological vector spaces

Throughout Part III, all vector spaces are over $\mathbb{R}$.

If E is a vector space, then a linear map $E \to \mathbb{R}$ is called a *linear functional*. The *algebraic dual space* E^* of E is the space of all linear functionals $E \to \mathbb{R}$.

A *topological vector space* is a vector space E equipped with a topology $\mathfrak{T}$ such that the map $E \times E \to E$, $(x, y) \mapsto x + y$ and the map $\mathbb{R} \times E \to E$, $(\lambda, x) \mapsto \lambda x$ are (jointly) continuous with respect to $\mathfrak{T}$. If E is a topological vector space, then its *topological dual space* E' is the space of all continuous linear functionals $E \to \mathbb{R}$. It is a subspace of the algebraic dual, but usually the two are different.

A topological vector space is *locally convex* if it has a neighbourhood base at 0 consisting of convex sets. Locally convex spaces are more well-behaved than general topological vector spaces, and almost all important spaces in functional analysis are locally convex. For more on locally convex spaces, the reader is referred to a graduate level textbook on functional analysis, for instance [Sch99, Rud91, Con07].

Dual pairs and weak topologies

Let E and F be vector spaces, and let $b : E \times F \to \mathbb{R}$ be a bilinear form. We say that a subset $N \subseteq F$ *separates points on* E (*via* b) if for every $x \in E$ there is some $y' \in N$

such that $b(x, y') \neq 0$. Likewise, a subset $M \subseteq E$ *separates points on* F (*via* b) if for every $y \in F$ there is some $x' \in M$ such that $b(x', y) \neq 0$. If F separates points on E and E separates points on F, then b is called a *dual pairing* (or *non-degenerate bilinear form*), and we denote it by the shorthand notation $\langle x, y\rangle := b(x, y)$. A *dual pair* is a tuple $(E, F, \langle \cdot\,, \cdot \rangle)$, where E and F are vector spaces and $\langle \cdot\,, \cdot \rangle : E \times F \to \mathbb{R}$ is a dual pairing. We usually use the shorthand notation $\langle E, F\rangle$ to denote the dual pair $(E, F, \langle \cdot\,, \cdot \rangle)$.

Let $\langle E, F\rangle$ be a dual pair. The $\sigma(E, F)$-*topology* on E is the initial topology induced by the family of linear functionals $\{x \mapsto \langle x, y\rangle \,:\, y \in F\}$, and the $\sigma(F, E)$-*topology* on F is the initial topology induced by the family of linear functionals $\{y \mapsto \langle x, y\rangle \,:\, x \in E\}$. If $F = E'$ is the topological dual of E, then the $\sigma(E, E')$-topology on E is called the *weak topology*, and the $\sigma(E', E)$-topology on E' is called the *weak-$*$ topology*. In this case, we denote the resulting topological vector spaces by E_w and E'_{w*}, respectively. Likewise, the weak closure of a subset $M \subseteq E$ will be denoted $\overline{M}^{\,w}$, and the weak-$*$ closure of a subset $N \subseteq E'$ will be denoted $\overline{N}^{\,w*}$.

Throughout Part III, we tacitly assume that all dual pairs consist of a topological vector space and its topological dual space (or algebraic dual space, if the primal space has no topology). Consequently, by a slight abuse of notation, we denote our dual pairs as $\langle E, E'\rangle$, $\langle F, F'\rangle$, etc., and we say that E *belongs to the dual pair* $\langle E, E'\rangle$. To keep the topological prerequisites to a minimum, we will forget about the original topology of E, and only remember the dual pair $\langle E, E'\rangle$ to which E belongs. When E has no topology, we tacitly assume that $E' := E^*$ is the algebraic dual.

Linear maps

If E and F are vector spaces, then the space of linear maps $E \to F$ is denoted by $\mathrm{L}(E, F)$. If E and F are topological vector spaces, then the space of *continuous* linear maps $E \to F$ is denoted by $\mathfrak{L}(E, F)$. If the (topological) duals separate points, then every continuous map $E \to F$ is also weakly continuous (see e.g. [Köt83, §20.4.(5)]), so we have

$$\mathfrak{L}(E, F) \subseteq \mathfrak{L}(E_w, F_w) \subseteq \mathrm{L}(E, F).$$

If E and F are vector spaces without topologies, then every linear map $T : E \to F$ is $\sigma(E, E^*)$-$\sigma(F, F^*)$-continuous (since $\psi \circ T$ is $\sigma(E, E^*)$-continuous for every $\psi \in F^*$), so we have

$$\mathfrak{L}(E_w, F_w) = \mathrm{L}(E, F) \qquad (\text{if } E' = E^*,\ F' = F^*).$$

The adjoint of a (continuous) linear map $T : E \to F$ is denoted $T^* : F^* \to E^*$ (algebraic adjoint) or $T' : F' \to E'$ (topological adjoint).

Bilinear maps

Let E, F, G be topological vector spaces. A bilinear map $b : E \times F \to G$ is (*jointly*) *continuous* if it is continuous with respect to the product topology on $E \times F$, and *separately continuous* if for all fixed $x_0 \in E$ and $y_0 \in F$ the maps $y \mapsto b(x_0, y)$ and $x \mapsto b(x, y_0)$ are continuous.

From left to right, let

$$\mathscr{Bil}(E \times F) \subseteq \mathfrak{Bil}(E \times F) \subseteq \mathrm{Bil}(E \times F)$$

denote the spaces of continuous, separately continuous, and all bilinear forms $E \times F \to \mathbb{R}$.[1]

For our purposes, the most important of these is $\mathfrak{Bil}(E \times F)$, the space of separately continuous bilinear forms. It will be used extensively in the study of the injective cone in Chapter 10.

Given a bilinear form $b \in \mathrm{Bil}(E \times F)$ and fixed vectors $x_0 \in E$, $y_0 \in F$, we let $b(x_0, \cdot\,) \in F^*$ and $b(\,\cdot\,, y_0) \in E^*$ denote the linear functionals

$$b(x_0, \cdot\,) := \big(y \mapsto b(x_0, y)\big);$$
$$b(\,\cdot\,, y_0) := \big(x \mapsto b(x, y_0)\big).$$

Using this notation, we see that b is separately continuous if and only if one has $b(x_0, \cdot\,) \in F'$ for all $x_0 \in E$ and $b(\,\cdot\,, y_0) \in E'$ for all $y_0 \in F$. In particular, it follows that $\mathfrak{Bil}(E \times F)$ does not depend on the topologies of E and F, but only on the dual pairs $\langle E, E'\rangle$, $\langle F, F'\rangle$. Likewise, it follows that $\mathfrak{Bil}(E \times F) = \mathrm{Bil}(E \times F)$ whenever $E' = E^*$ and $F' = F^*$.

It follows from [Köt79, §40.1.(2')] and the preceding remarks that the maps $b \mapsto (x \mapsto b(x, \cdot\,))$ and $b \mapsto (y \mapsto b(\,\cdot\,, y))$ define linear isomorphisms

$$\mathfrak{Bil}(E \times F) = \mathfrak{Bil}(E_w \times F_w) \cong \mathfrak{L}(E_w, F'_{w*}) \cong \mathfrak{L}(F_w, E'_{w*}).$$

(The isomorphism $\mathfrak{L}(E_w, F'_{w*}) \cong \mathfrak{L}(F_w, E'_{w*})$ is simply $T \mapsto T'$.)

Tensor products

We assume the reader to be familiar with the basics of the (algebraic) theory of tensor products. We will need very little on the side of topological tensor products (but many results in Part III are inspired by the theory of normed tensor products).

For clarity, we shall occasionally use the following notation: if E and F are vector spaces and $M \subseteq E$ and $N \subseteq F$ are subsets, then we define the "set-wise" tensor product

$$M \otimes_s N := \{x \otimes y \,:\, x \in M,\ y \in N\} \subseteq E \otimes F.$$

8.2 Subspaces, quotients, and tensor products of dual pairs

Many of the properties of a convex cone E_+ in topological vector space E depend only on the geometry of E_+ and on the dual pair $\langle E, E'\rangle$, not on the precise topology of E. In particular, we don't need to know the exact topology of $E \otimes F$, because for our purposes it suffices to know what its dual space is. This enables us to ignore topological issues in the tensor product, thereby circumventing the notoriously complicated theory

[1] Note: with this notation it is possible to confuse $\mathscr{Bil}(E \times F)$ with $\mathfrak{Bil}(E \times F)$, but notation like this appears to be at least moderately common (e.g. [Sch99, p. 91], [Köt79, p. 154]).

of locally convex tensor products. Instead, we formulate our results for a wide range of *reasonable* duals of $E \otimes F$ (see below).

Throughout Part III, we encode the "input spaces" E and F and the "output space" $E \otimes F$ by the dual pairs to which they belong; that is, by only remembering what the appropriate (algebraic or topological) dual space is, without remembering the exact topology. In this section, we briefly discuss how to handle subspaces, quotients, and tensor products of dual pairs.

Questions about the projective/injective cone that depend not only on the dual pair, but also on a specific topology on $E \otimes F$, will not be treated in this dissertation. In particular, for questions about normality of the projective/injective cone, we refer the reader to [PS69].

Remark 8.1. Because we choose to forget about the topology of E and only formulate results in terms of the dual pair $\langle E, E' \rangle$, we occasionally have to make use of the weak topology. In particular, we often refer to the *weak closure* of a convex cone and to *weakly closed* subspaces. We should point out that the adjective "weak" can be omitted here if E is a locally convex space, because in this setting the weak and original closure of a convex set (in particular, a convex cone or a subspace) coincide, by [Rud91, Theorem 3.12].

If E is a topological vector space which is not locally convex, then the adjective "weak" cannot be omitted.

Subspaces

If $\langle E, E' \rangle$ is a dual pair and if $I \subseteq E$ is a subspace, then we will understand I to belong to the dual pair $\langle I, E'/I^\perp \rangle$.

We show that this is usually, but not always, the natural dual pair for I. To that end, assume that E a topological vector space, E' is its (topological) dual, and I carries the subspace topology. Let $T : I \hookrightarrow E$ denote the inclusion and $T' : E' \to I'$ its adjoint.

If E is locally convex, then every continuous linear functional on I can be extended to E, so T' is surjective. Clearly $\ker(T') = I^\perp$, so T' restricts to a linear isomorphism $E'/I^\perp \to I'$. Furthermore, the relative $\sigma(E, E')$-topology on I coincides with the $\sigma(I, E'/I^\perp)$-topology (even if I is not closed), so we may unambiguously refer to this as the weak topology on I. On the other hand, the $\sigma(E'/I^\perp, I)$-topology on $E'/I^\perp = I'$ coincides with the quotient topology $E'_{w*}/I^\perp$ if and only if I is closed (see e.g. [Sch99, §IV.4.1, Corollary 1]).

If E is not locally convex, then I may have continuous linear functionals that cannot be extended. In this case one still has $\ker(T') = I^\perp$, but T' is not surjective, so $I' \neq E'/I^\perp$. Nevertheless, $E'/I^\perp$ is the dual of I with respect to the relative $\sigma(E, E')$-topology on I.

Quotients

If E is a topological vector space and if $I \subseteq E$ is a closed subspace, then E/I is a Hausdorff topological vector space. Every continuous linear functional $E/I \to \mathbb{R}$

extends to a continuous linear functional $E \to \mathbb{R}$ that vanishes on I. Conversely, if $\varphi : E \to \mathbb{R}$ is a continuous linear functional that vanishes on I, then φ factors through E/I, by the universal property of quotients. Therefore: $(E/I)' \cong I^\perp$ as vector spaces.

Thus, if $\langle E, E' \rangle$ is a dual pair and if $I \subseteq E$ is a weakly closed subspace, then we can understand E/I to belong to the dual pair $\langle E/I, I^\perp \rangle$. The quotient topology on E_w/I coincides with the $\sigma(E/I, I^\perp)$-topology, and the subspace topology on $I^\perp \subseteq E'_{w*}$ coincides with the $\sigma(I^\perp, E/I)$-topology (see e.g. [Sch99, §IV.4.1, Corollary 1]), so we may unambiguously refer to these as the weak topology on E/I and the weak-$*$ topology on $I^\perp$, respectively.

The only downside to this approach is that we cannot "see" all quotients of E. If E is locally convex, then every closed subspace is also weakly closed, but this is not true for general topological vector spaces (see e.g. [Kal78]). However, if I is closed but not weakly closed, then the quotient E/I is Hausdorff, but its topological dual $(E/I)' = I^\perp$ does not separate points. Throughout this dissertation, we assume that all duals separate points, so we only consider quotients E/I where I is weakly closed.

Tensor products

Let $\langle E, E' \rangle$ and $\langle F, F' \rangle$ be dual pairs. Recall from §8.1 that the space $\mathfrak{Bil}(E \times F)$ of separately continuous bilinear forms $E \times F \to \mathbb{R}$ can be defined without specifying topologies on E and F, since this space depends only on the dual pairs $\langle E, E' \rangle$ and $\langle F, F' \rangle$. Since the algebraic dual of $E \otimes F$ is isomorphic with $\mathrm{Bil}(E \times F)$, we can identify $\mathfrak{Bil}(E \times F)$ with a subspace of $(E \otimes F)^*$. We say that a subspace $G \subseteq (E \otimes F)^*$ is a *reasonable dual of* $E \otimes F$ (with respect to the dual pairs $\langle E, E' \rangle$, $\langle F, F' \rangle$) if

$$E' \otimes F' \subseteq G \subseteq \mathfrak{Bil}(E \times F).$$

This definition will allow us to treat (duality of) convex cones in topological tensor products without having to deal with the specifics of topological tensor products.

We show that this definition covers all important cases. First, if E, F are locally convex and $E \otimes F$ carries a *compatible* topology α (in the sense of Grothendieck [Gro55, p. 89]; see also [Köt79, §44.1]), then we claim that the topological dual $(E \otimes_\alpha F)'$ is a reasonable dual of $E \otimes F$. Indeed, one of the requirements for α to be compatible is $E' \otimes F' \subseteq (E \otimes_\alpha F)'$. Moreover, every compatible topology is coarser than the inductive topology, whose dual is $\mathfrak{Bil}(E \times F)$ (see e.g. [Köt79, §44.1.(5)]), so one has $(E \otimes_\alpha F)' \subseteq \mathfrak{Bil}(E \times F)$. This shows that $(E \otimes_\alpha F)'$ is a reasonable dual.

Second, if E and F originate from spaces without topologies, then we understand these to belong to the dual pairs $\langle E, E^* \rangle$, $\langle F, F^* \rangle$. In this case we have $\mathfrak{Bil}(E \times F) = \mathrm{Bil}(E \times F)$ (see §8.1), so we find that $(E \otimes F)^* = \mathrm{Bil}(E \times F)$ is a reasonable dual of $E \otimes F$. This is useful when applying topological results in the non-topological setting (for instance, see Corollary 9.4).

8.3 Ordered vector spaces

Convex cones and their duals

Let E be a (real) vector space. A *convex cone*[2] is a non-empty subset $\mathcal{K} \subseteq E$ satisfying $\mathcal{K} + \mathcal{K} \subseteq \mathcal{K}$ and $\lambda\mathcal{K} \subseteq \mathcal{K}$ for all $\lambda \in \mathbb{R}_{\geq 0}$. If $\mathcal{K}$ is a convex cone, then $\mathrm{lin}(\mathcal{K}) := \mathcal{K} \cap -\mathcal{K}$ is a linear subspace of E, called the *lineality space* of $\mathcal{K}$. We say that $\mathcal{K}$ is *proper*[3] if $\mathrm{lin}(\mathcal{K}) = \{0\}$, and *generating* if $\mathcal{K} - \mathcal{K} = E$.

If $\mathcal{K} \subseteq E$ is a convex cone, then its *algebraic dual cone* $\mathcal{K}^*$ is the set of all positive linear functionals:

$$\mathcal{K}^* := \big\{\varphi \in E^* \,:\, \varphi(x) \geq 0 \text{ for all } x \in \mathcal{K}\big\}.$$

If $\langle E, E'\rangle$ is a dual pair, then we define $\mathcal{K}' := \mathcal{K}^* \cap E'$ (the *dual cone* for the dual pair $\langle E, E'\rangle$). The dual cone of $\mathcal{K}' \subseteq E'$ with respect to the dual pair $\langle E', E\rangle$ is the *bipolar cone*

$$\mathcal{K}'' := \big\{x \in E \,:\, \langle x, \varphi\rangle \geq 0 \text{ for all } \varphi \in \mathcal{K}'\big\} = (\mathcal{K}')'.^4$$

Using the (one-sided) bipolar theorem, one easily shows that $\mathcal{K}'' = \overline{\mathcal{K}}^{w}$. It follows that ${}^{\perp}(\mathcal{K}') = \mathcal{K}'' \cap -\mathcal{K}'' = \mathrm{lin}(\overline{\mathcal{K}}^{w})$. In particular, $\overline{\mathcal{K}}^{w}$ is a proper cone if and only if $\mathcal{K}'$ separates the points of E. If this is the case, then we say that $\mathcal{K}$ is *semisimple*. (For an equivalent definition of semisimplicity in terms of representations, see [Dob20a].)

Ordered vector spaces

Let E be a vector space. A *vector preorder* is a preorder $\leq$ on E such that for all $x, y, z \in E$ and $\lambda \in \mathbb{R}_{>0}$ one has $x \leq y$ if and only if $x + z \leq y + z$ if and only if $\lambda x \leq \lambda y$.

There is a natural bijective correspondence between vector preorders on E and convex cones in E, which identifies the preorder $\leq$ with the convex cone $E_+ := \{x \in E \,:\, x \geq 0\}$ of positive elements of E. In the reverse direction, a convex cone $\mathcal{K} \subseteq E$ is identified with the vector preorder $\leq_{\mathcal{K}}$ given by $x \leq_{\mathcal{K}} y$ if and only if $y - x \in \mathcal{K}$ (for all $x, y \in E$).

A *preordered vector space* is a tuple (E, E_+) where E is a vector space and $E_+ \subseteq E$ is a convex cone. We understand E to be preordered by the vector preorder associated with E_+. Likewise, a *preordered topological vector space* is a tuple (E, E_+), where E is a topological vector space and $E_+ \subseteq E$ is a convex cone. Note that we do not assume any kind of compatibility between the topology and the cone E_+.

The positive cone E_+ of a preordered (topological) vector space E is proper if and only if the associated vector preorder is antisymmetric (so it is a partial order). If this is the case, then (E, E_+) is called an *ordered* (*topological*) *vector space*.

---

[2] A note about terminology: some authors call this a *wedge*, and reserve the term *cone* for what we call a *proper cone* (e.g. [Day62, Per67, AT07]).

[3] Some authors call this *pointed* or *salient*.

[4] There is some chance of confusion here, because $\mathcal{K}''$ could also refer to the positive cone of the second dual E'' of E. To avoid this confusion, and in light of the bipolar theorem, we will henceforth refer to the bipolar cone as $\overline{\mathcal{K}}^{w}$ instead of $\mathcal{K}''$.

Whenever we have a (topological) vector space E and a convex cone $E_+ \subseteq E$, we will implicitly assume that E is a preordered (topological) vector space with positive cone E_+. Furthermore, the preorder of E will be denoted by $\leq$.

Positive linear maps

Let (E, E_+) and (F, F_+) be preordered vector spaces. We say that a linear map $T \in \mathrm{L}(E, F)$ is *positive* if $T[E_+] \subseteq F_+$, a *pullback* (or *bipositive operator*) if $E_+ = T^{-1}[F_+]$, and a *pushforward* if $T[E_+] = F_+$.

Furthermore, if E and F belong to dual pairs $\langle E, E' \rangle$ and $\langle F, F' \rangle$, then we say that an operator $T \in \mathfrak{L}(E_w, F_w)$ is *approximately positive* if $T[\overline{E_+}^w] \subseteq \overline{F_+}^w$, an *approximate pullback* (or *approximately bipositive*) if $\overline{E_+}^w = T^{-1}[\overline{F_+}^w]$, and an *approximate pushforward* if $\overline{T[E_+]}^w = \overline{F_+}^w$. A continuous positive map (resp. pushforward) is also approximately positive (resp. an approximate pushforward), but a pullback is not necessarily an approximate pullback. (Concrete example: let $F = \mathbb{R}^2$ with $F_+ = \{(x, y) : x > 0\} \cup \{(0, 0)\}$, let $E := \mathrm{span}\{(0, 1)\} \subseteq F$ with $F_+ := F_+ \cap E$, and let T be the inclusion $E \hookrightarrow F$.)

These approximate type operators are not particularly natural from the perspective of ordered vector spaces, but they come into play as soon as one starts to make use of duality. Given $T \in \mathfrak{L}(E_w, F_w)$, it is not hard to show that the adjoint $T' \in \mathfrak{L}(F'_{w*}, E'_{w*})$ is positive if and only if T is approximately positive. In addition, using that $(T[C])^\circ = (T')^{-1}[C^\circ]$ (e.g. [Sch99, Proposition IV.2.3(a)]), it is easy to show that T is an approximate pullback if and only if T' is a weak-$*$ approximate pushforward, and vice versa. This is no longer true if the adjective "approximate" is omitted.

We shall treat pullbacks and pushforwards as the natural ordered analogues of metric injections (isometries) and metric surjections (quotients); see Table 8.2. As soon as duality comes into play, it will be helpful to pass to the corresponding approximate versions. In particular, we show that the injective cone preserves approximate pullbacks, but not all pullbacks.

Note that every linear map $E \to F$ can be made a pullback/pushforward by choosing appropriate cones. In particular, a pullback is not necessarily injective, and a pushforward is not necessarily surjective. However, if E_+ is a proper cone, then every pullback $T : E \to F$ is injective (since $\ker(T) \subseteq T^{-1}[F_+] = E_+$), and if F_+ is generating then every pushforward $E \to F$ is surjective.

Table 8.2: Ordered analogues of common concepts in the normed theory.

Normed theory	Ordered theory
Continuous operator	Positive operator
Metric injection (isometry)	Pullback (bipositive operator)
Metric surjection (quotient)	Pushforward (quotient)
Projection (complemented subspace)	Positive projection (order retract)

Retracts

Let (E, E_+) be a preordered vector space. A subspace $F \subseteq E$ is an *order retract* if there exists a positive projection $E \to F$. If E is furthermore a topological vector space, then we say that F is a *topological order retract* if there exists a *continuous* positive projection $E \to F$.

For simplicity, we shall speak of *retracts* and *top-retracts*, as there is minimal chance of confusion with other types of retracts (e.g. from topology).

Note that a retract provides at the same time an injective pullback (i.e. bipositive map) $F \hookrightarrow E$ and a surjective pushforward ("quotient") $E \twoheadrightarrow F$. We will show that, although the projective tensor product does not preserve bipositive maps and the injective tensor product does not preserve quotients, retracts are sufficiently rigid to be preserved by both.

To illustrate their place in the theory, note that every top-retract is a complemented subspace (after all, it admits a continuous projection[5]). If $E_+ = \{0\}$, then the top-retracts are precisely the complemented subspaces.

As far as we know, order retracts are not a very common notion, and have not received much attention. However, some special cases already play a role in the theory, such as *projection bands* in Riesz spaces (see e.g. [Zaa97, §11]) and *projectionally exposed faces* in finite-dimensional cones (see e.g. [BLP87, ST90]).

Positive bilinear maps

If (E, E_+), (F, F_+), (G, G_+) are preordered vector spaces, then a bilinear map $b : E \times F \to G$ is called *positive* if $b(E_+, F_+) \subseteq G_+$.

In terms of the isomorphism $\mathfrak{Bil}(E_w \times F_w) \cong \mathfrak{L}(E_w, F'_{w*})$ (see §8.1), we note that a bilinear form $b \in \mathfrak{Bil}(E_w \times F_w)$ is positive if and only if $b(x, \cdot\,)$ defines a positive linear functional on F for every $x \in E_+$, or equivalently, if and only if the corresponding map $E_w \to F'_{w*}$ is positive. Thus, contrary to the topological setting, there is no difference between positive and "separately positive" bilinear forms.

Faces and extremal rays

Let E be a vector space and let $E_+ \subseteq E$ be a convex cone. A *face* (or *extremal set*) of E_+ is a (possibly empty) convex subset $M \subseteq E_+$ such that, if M contains a point in the relative interior of a line segment in E_+, then M also contains the endpoints of that segment. If φ is a continuous positive linear functional, then $\ker(\varphi) \cap E_+$ is a face. Faces of this type are called *exposed*.

Every convex cone has a unique minimal non-empty face (the lineality space $\mathrm{lin}(E_+)$, contained in every face) and a unique maximal face (the cone itself, containing every face). Note that E_+ is a proper cone if and only if $\{0\}$ is a face.

Let $x_0 \in E_+ \setminus \{0\}$. If $M := \{\lambda x_0 : \lambda \in \mathbb{R}_{\geq 0}\}$ is a face, then we say that x_0 is an *extremal direction*, and M is an *extremal ray*. If x_0 is an extremal direction, then so

[5]Some authors require a complemented subspace to be closed, but this is automatic: if $P : E \to E$ is a continuous projection with range F, then $F = \ker(\mathrm{id}_E - P)$, so F is closed.

is μx_0 for every $\mu > 0$. We let $\operatorname{rext}(E_+) \subseteq E_+ \setminus \{0\}$ denote the set of all extremal directions of E_+.

If $M \subseteq E_+$ is a non-empty subset, then $E'_+ \cap M^\perp$ defines a face of E'_+. Faces of this type are called *dual faces*. In the finite-dimensional case, dual faces are precisely the exposed faces, but this is not true in locally convex spaces. For more on dual and exposed faces, see Appendix A.3.

Order ideals

Let (E, E_+) be a preordered vector space. A subspace $I \subseteq E$ is called an *order ideal* if the pushforward of E_+ along the quotient map $E \to E/I$ is a proper cone. If no ambiguity can arise (i.e. if the space does not carry a multiplication), then we call I simply an *ideal*.

A subspace $I \subseteq E$ is an ideal if and only if $I \cap E_+$ is a face of E_+ (see Proposition A.2). Conversely, if $M \subseteq E_+$ is a face, then $\operatorname{span}(M)$ is an ideal satisfying $\operatorname{span}(M) \cap E_+ = M$ (see Proposition A.3). Thus, $I \mapsto I_+$ defines a many-to-one correspondence between ideals and faces. We shall draw heavily upon this correspondence.

If $\mathcal{K} \subseteq E_+$ is a subcone, then every ideal $I \subseteq E$ with respect to E_+ is also an ideal with respect to $\mathcal{K}$. More generally, if $T : E \to F$ is a positive linear map and if $J \subseteq F$ is an ideal, then $T^{-1}[J] \subseteq E$ is also an ideal (see Proposition A.3). In particular, if F_+ is a proper cone, then $\{0\} \subseteq F_+$ is a face, so $\ker(T) \cap E_+$ is a face of E_+. It can be shown that all faces can be written in this form (see Proposition A.4(b)).

We will show in Corollary A.12 that the maximal order ideals are precisely the kernels of non-zero positive linear functionals, or in other words, the supporting hyperplanes of E_+. In particular, not every maximal ideal in a preordered topological vector space is closed. (Example: the kernel of a discontinuous positive linear functional.)

For more about ideals and faces, see Appendix A and [Bon54].

Chapter 9

The projective cone

In this chapter, we carry out an in-depth study of the properties of the projective cone. This cone does not depend on any topological data, so we will mostly ignore topological issues in this chapter. Some questions about the closure of the projective cone will briefly be discussed in Chapter 11.

This chapter is based on Chapter 3 of [Dob20b].

Introduction

Let E, F be (real) vector spaces and let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones. The simplest way to define a cone in $E \otimes F$ is to consider the *projective cone*

$$E_+ \otimes^\pi F_+ := \left\{ \sum_{i=1}^k x_i \otimes y_i \ : \ k \in \mathbb{N},\ x_1, \dots, x_k \in E_+,\ y_1, \dots, y_k \in F_+ \right\}.$$

If E, F are locally convex and if α is a compatible locally convex topology on $E \otimes F$, then we denote by $E_+ \otimes_\alpha^\pi F_+$ and $E_+ \tilde{\otimes}_\alpha^\pi F_+$ the same cone, but embedded in the topological vector spaces $E \otimes_\alpha F$ and $E \tilde{\otimes}_\alpha F$, respectively. (The topology is denoted in the subscript; the cone in the superscript.)

It is easy to see that $E_+ \otimes^\pi F_+$ is indeed a (convex) cone. This cone has received a lot of attention in the literature; see for instance [Mer64, PS69, GL88, GK10, Wor19].

In the subsequent sections, we will study the basic properties of the projective cone. We point out a characteristic property of the projective cone (§9.1), study its mapping properties (§9.2), prove precise necessary and sufficient conditions for $E_+ \otimes^\pi F_+$ to be proper (§9.3), and show that the projective tensor product of two faces is again a face (§9.4, §9.5). Finally, as an application of the results from this section, we prove that the tensor product of absolutely convex sets also preserves faces (§9.6).

9.1 The characteristic property of the projective cone

Let E, F, G be vector spaces equipped with convex cones $E_+ \subseteq E$, $F_+ \subseteq F$, $G_+ \subseteq G$. There is a natural isomorphism $\mathrm{Bil}(E \times F, G) \cong L(E \otimes F, G)$, which identifies a bilinear map $\Phi : E \times F \to G$ with its linearization $\Phi^L : E \otimes F \to G$, $\Phi^L(\sum_{i=1}^k x_i \otimes y_i) = \sum_{i=1}^k \Phi(x_i, y_i)$.

Proposition 9.1. *If $E \otimes F$ is equipped with the projective cone $E_+ \otimes^\pi F_+$, then a linear map $\Phi^L : E \otimes F \to G$ is positive if and only if its corresponding bilinear map $\Phi : E \times F \to G$ is positive.*

Proof. A bilinear map $\Phi : E \times F \to G$ is positive if and only if $\Phi^L(x \otimes y) = \Phi(x, y) \geq 0$ for all $x \in E_+$, $y \in F_+$. On the other hand, since $E_+ \otimes^\pi F_+$ is generated by $E_+ \otimes F_+$, we also find that a linear map $\Phi^L : E \otimes F \to G$ is positive if and only if $\Phi^L(x \otimes y) \geq 0$ for all $x \in E_+$, $y \in F_+$. ■

This is the ordered analogue of the characteristic property of the projective topology. It follows that

$$(E_+ \otimes^\pi F_+)^* = \mathrm{Bil}(E \times F)_+ \qquad (E, F \text{ vector spaces});$$
$$(E_+ \otimes_\pi^\pi F_+)' = (E_+ \tilde{\otimes}_\pi^\pi F_+)' = \mathscr{B}il(E \times F)_+ \qquad (E, F \text{ locally convex}).$$

9.2 Mapping properties of the projective cone

The projective norm preserves continuous linear maps, quotients, and complemented subspaces (see e.g. [DF93, Propositions 3.2, 3.8, and 3.9(1)], or [Köt79, §41.5] for the more general locally convex setting). The projective cone has analogous mapping properties.

Proposition 9.2. *Let $T \in \mathrm{L}(E, G)$ and $S \in \mathrm{L}(F, H)$.*

(a) If $T[E_+] \subseteq G_+$ and $S[F_+] \subseteq H_+$, then $(T \otimes S)[E_+ \otimes^\pi F_+] \subseteq G_+ \otimes^\pi H_+$.

(b) If $T[E_+] = G_+$ and $S[F_+] = H_+$, then $(T \otimes S)[E_+ \otimes^\pi F_+] = G_+ \otimes^\pi H_+$.

(c) If (E, E_+) and (F, F_+) are retracts of (G, G_+) and (H, H_+), respectively, then $(E \otimes F, E_+ \otimes^\pi F_+)$ is a retract of $(G \otimes H, E_+ \otimes^\pi F_+)$.

In summary: the projective cone preserves positive linear maps, pushforwards, and retracts.

It follows immediately that the same statements hold for maps between the completions (in the locally convex case), for the projective cone is contained in the algebraic tensor product.

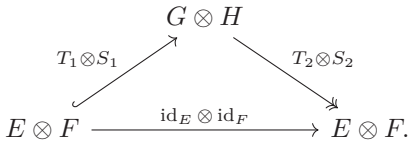
Proof.

(a) Let $z \in E_+ \otimes^\pi F_+$ be given, and write $z = \sum_{i=1}^k x_i \otimes y_i$ with $x_1, \ldots, x_k \in E_+$, $y_1, \ldots, y_k \in F_+$. Then we have $(T \otimes S)(z) = \sum_{i=1}^k T(x_i) \otimes S(y_i) \in G_+ \otimes^\pi H_+$, since $T(x_1), \ldots, T(x_k) \in G_+$, $S(y_1), \ldots, S(y_k) \in H_+$.

(b) By (a), $T \otimes S$ is positive. Now let $u \in G_+ \otimes^\pi H_+$ be given, and write $u = \sum_{i=1}^k v_i \otimes w_i$ with $v_1, \ldots, v_k \in G_+$ and $w_1, \ldots, w_k \in H_+$. By assumption there are $x_1, \ldots, x_k \in E_+$, $y_1, \ldots, y_k \in F_+$ such that $v_i = T(x_i)$ and $w_i = S(y_i)$, for all i. Consequently, we have $z := \sum_{i=1}^k x_i \otimes y_i \in E_+ \otimes^\pi F_+$, and $u = (T \otimes S)(z)$.

(c) There are positive linear maps T_1, T_2, S_1, S_2 so that the following two diagrams commute:

Consequently, the following diagram commutes:



By (a), the maps in the preceding diagram are all positive for the projective cone, so it follows that $(E \otimes F, E_+ \otimes^\pi F_+)$ is a retract of $(G \otimes H, G_+ \otimes^\pi H_+)$. ■

Next, we prove that the projective tensor product also preserves approximate pushforwards: if T and S are maps whose adjoints are bipositive, then the same is true of $T \otimes S$.

Lemma 9.3. *Let $\langle E, E'\rangle$, $\langle F, F'\rangle$, $\langle G, G'\rangle$, $\langle H, H'\rangle$ be dual pairs, and let E_+, F_+, G_+, H_+ be convex cones in the primal spaces. If $T \in \mathfrak{L}(E_w, G_w)$ and $S \in \mathfrak{L}(F_w, H_w)$ are approximate pushforwards, then the map $(T \otimes S)' : \mathfrak{Bil}(G_w \times H_w) \to \mathfrak{Bil}(E_w \times F_w)$, $((T \otimes S)'b)(x, y) = b(Tx, Sy)$ is bipositive.*

Here $(T \otimes S)'$ denotes the adjoint of $T \otimes S : E \otimes F \to G \otimes H$, assuming that $E \otimes F$ and $G \otimes H$ are equipped with the largest reasonable duals (see §8.2).

Proof. Note that $(T \otimes S)'b$ is a positive bilinear functional on $E \times F$ if and only if b is positive on $T[E_+] \times S[F_+]$, so if b is separately weakly continuous, then this is the case if and only if b is positive on $\overline{T[E_+]}^w \times \overline{S[F_+]}^w$. (First use weak continuity in the second variable to pass from $T[E_+] \times S[F_+]$ to $T[E_+] \times \overline{S[F_+]}^w$, then use weak continuity in the first variable to proceed to $\overline{T[E_+]}^w \times \overline{S[F_+]}^w$.) Analogously, b itself is a positive bilinear functional on $G \times H$ if and only if b is positive on $\overline{G_+}^w \times \overline{H_+}^w$. By assumption, we have $\overline{T[E_+]}^w = \overline{G_+}^w$ and $\overline{S[F_+]}^w = \overline{H_+}^w$, so it follows that b is positive if and only if $(T \otimes S)'b$ is positive. ■

The preceding lemma has immediate applications to algebraic tensor products (of vector spaces without topologies) and to completed locally convex topologies. It will also be used to prove one of the fundamental properties of the injective cone (see Lemma 10.15(b)).

Corollary 9.4. *Let E, F, G, H be preordered vector spaces, and let $T \in \mathrm{L}(E, G)$, $S \in \mathrm{L}(F, H)$ be linear maps such that T^* and S^* are bipositive. Then $(T \otimes S)^*$ is bipositive with respect to the dual cones $(E_+ \otimes^\pi F_+)^* \subseteq (E \otimes F)^*, (G_+ \otimes^\pi H_+)^* \subseteq (G \otimes H)^*$.*

Proof. If we understand the primal spaces to belong to the dual pairs $\langle E, E^*\rangle$, ..., $\langle H, H^*\rangle$, then every linear map is weakly continuous. Furthermore, $(E \otimes F)^* = \mathrm{Bil}(E \times F) = \mathfrak{Bil}(E_w \times F_w)$, and the positive cone of $\mathfrak{Bil}(E_w \times F_w)_+$ coincides with the dual cone $(E_+ \otimes^\pi F_+)^* \subseteq (E \otimes F)^*$, by Proposition 9.1. Hence the result is a special case of Lemma 9.3. ■

Corollary 9.5. *Let E, F, G, H be locally convex preordered topological vector spaces and let $T \in \mathfrak{L}(E, G)$ and $S \in \mathfrak{L}(F, H)$ be approximate pushforwards. If $T \otimes_{\alpha\to\beta} S : E \otimes_\alpha F \to G \otimes_\beta H$ is continuous (α and β compatible locally convex topologies), then $T \otimes_{\alpha\to\beta} S$ and $T \tilde{\otimes}_{\alpha\to\beta} S$ are approximate pushforwards.*

Proof. Every continuous linear map is also weakly continuous (see [Köt83, §20.4.(5)]), so we have $T \in \mathfrak{L}(E_w, G_w)$ and $S \in \mathfrak{L}(F_w, H_w)$. Furthermore, since α and β are compatible topologies, we have $(E \otimes_\alpha F)' \subseteq \mathfrak{Bil}(E \times F) = \mathfrak{Bil}(E_w \times F_w)$ and $(G \otimes_\beta H)' \subseteq \mathfrak{Bil}(F \times H) = \mathfrak{Bil}(F_w \times H_w)$. It follows that $(T \otimes_{\alpha\to\beta} S)'$ is a restriction of the map $(T \otimes S)'$ from Lemma 9.3, and therefore it is also bipositive. For the completion, note that $(T \tilde{\otimes}_{\alpha\to\beta} S)' = (T \otimes_{\alpha\to\beta} S)'$. ■

Interestingly, Corollary 9.4 uses topological techniques to prove a purely algebraic result. We don't know a purely algebraic proof of Corollary 9.4.

Finally, we show that the projective tensor product does not preserve bipositive maps, even if all spaces are finite-dimensional and all cones are closed and generating (Example 9.6), or even closed, generating and proper (Example 9.7).

Example 9.6. As a very simple example, let $F = G = \mathbb{R}^2$ with $F_+ = \mathbb{R}^2$ and $G_+ = \mathbb{R}^2_{\geq 0}$. Furthermore, let $E = \mathrm{span}\{(1, -1)\} \subseteq G$, and write $E_+ := E \cap G_+ = \{0\}$. Then the inclusion $T : E \hookrightarrow G$ is bipositive, but $E_+ \otimes^\pi F_+ = \{0\}$ whereas $G_+ \otimes^\pi F_+ = G \otimes F$. Since $E \otimes F \neq \{0\}$, we have $(G_+ \otimes^\pi F_+) \cap (E \otimes F) \neq E_+ \otimes^\pi F_+$, which shows that $T \otimes \mathrm{id}_F$ is not bipositive. △

Example 9.7 (Compare [Dob22, Situation 4])**.** For a more advanced example, let E be a finite-dimensional space equipped with a proper, generating, polyhedral cone E_+ which is *not* a simplex cone. Choose $\varphi_1, \ldots, \varphi_m \in E^*$ such that $E_+ = \bigcap_{i=1}^m \{x \in E : \varphi_i(x) \geq 0\}$, and let $\mathbb{R}^m$ be equipped with the standard cone $\mathbb{R}^m_{\geq 0}$. Then the map $T : E \to \mathbb{R}^m$, $x \mapsto (\varphi_1(x), \ldots, \varphi_m(x))$ is bipositive.

Since E_+ is not a simplex cone, it follows from [BL75, Proposition 3.1] (see also Theorem 13.2 below) that $E_+ \otimes^\pi E_+^* \neq E_+ \otimes^\varepsilon E_+^*$. On the other hand, it is well-known that $\mathbb{R}^m_{\geq 0} \otimes^\pi E_+^* = \mathbb{R}^m_{\geq 0} \otimes^\varepsilon E_+^*$, and it follows from Theorem 10.16(b) below that $T \otimes \mathrm{id}_{E^*}$ is bipositive for the injective cone. Therefore:

$$(T \otimes \mathrm{id}_{E^*})^{-1}[\mathbb{R}^m_{\geq 0} \otimes^\pi E_+^*] = (T \otimes \mathrm{id}_{E^*})^{-1}[\mathbb{R}^m_{\geq 0} \otimes^\varepsilon E_+^*] = E_+ \otimes^\varepsilon E_+^* \neq E_+ \otimes^\pi E_+^*.$$

This shows that $T \otimes \mathrm{id}_{E^*}$ is not bipositive for the projective cone.

Note that all cones in this example are polyhedral, and therefore closed. In particular, the situation is not resolved by taking closures. △

The finite-dimensional techniques used in Example 9.7 will be discussed in more detail in Chapter 12 and Chapter 13.

Despite the preceding counterexamples, bipositivity can be preserved under certain additional conditions. First, if $E \subseteq G$ and $F \subseteq H$ are retracts, then $E \otimes F \subseteq G \otimes H$ is also a retract (by Proposition 9.2(c)), so in particular the inclusion $E \otimes F \hookrightarrow G \otimes H$ is bipositive. Furthermore, we prove in Proposition 9.21 that the projective cone also preserves ideals of proper cones bipositively.

9.3 When is the projective cone proper?

There is a simple necessary and sufficient condition for $E_+ \otimes^\pi F_+$ to be proper, which we prove in Theorem 9.10 below. This result was first proved (in three different ways) by Dermenjian and Saint-Raymond [DS70], and recently rediscovered by Wortel [Wor19]. (The original proof seems to have been forgotten, and before Wortel only special cases were known in the literature.) The proof given here is different from each of the existing proofs. Further methods of proof will be discussed in Remark 9.12.

We proceed via reduction to the finite-dimensional case, using the following lemmas.

Lemma 9.8. *A convex cone $E_+ \subseteq E$ is generating if and only if its algebraic dual cone E_+^* is proper.*

Proof. Note that E_+^* is *not* proper if and only if there is some $\varphi \in E^* \setminus \{0\}$ such that both φ and $-\varphi$ are positive linear functionals, or equivalently, $\varphi(x) = 0$ for all $x \in E_+$. This is in turn equivalent to E_+ being contained in a (linear) hyperplane, which happens if and only if E_+ is *not* generating. ∎

Corollary 9.9. *If E is finite-dimensional, then a closed convex cone $E_+ \subseteq E$ is proper if and only if its dual cone E_+^* is generating.*

Proof. Set $F := E^*$ and $F_+ := E_+^*$. Under the canonical isomorphism $E \cong E^{**}$, we have $F_+^* = E_+$, by the bipolar theorem (here we use that E_+ is closed). The result follows from Lemma 9.8, applied to the cone $F_+ \subseteq F$. ∎

We are now ready to state and prove the main result of this section.

Theorem 9.10 (cf. [DS70])**.** *Let E and F be vector spaces with convex cones $E_+ \subseteq E$, $F_+ \subseteq F$. Then the projective cone $E_+ \otimes^\pi F_+$ is proper if and only if $E_+ = \{0\}$, or $F_+ = \{0\}$, or both E_+ and F_+ are proper.*

Proof. Suppose first that $E_+, F_+ \neq \{0\}$ and E_+ is *not* proper. Then we may choose $x \in E \setminus \{0\}$ such that $x, -x \in E_+$, and $y \in F_+ \setminus \{0\}$. Both $x \otimes y$ and $-x \otimes y$ belong to $E_+ \otimes^\pi F_+$, but we have $x \otimes y \neq 0$, so we see that $E_+ \otimes^\pi F_+$ is not a proper cone.

For the converse, if $E_+ = \{0\}$, then $E_+ \otimes^\pi F_+ = \{0\}$ regardless of any properties of F_+ (and similarly if $F_+ = \{0\}$). So assume now that both E_+ and F_+ are proper (not necessarily $\neq \{0\}$). Let $z \in E \otimes F$ be given such that $z, -z \in E_+ \otimes^\pi F_+$. Then we may choose integers $n \geq k \geq 0$ and vectors $x_1, \ldots, x_n \in E_+$, $y_1, \ldots, y_n \in F_+$ such that $z = \sum_{i=1}^k x_i \otimes y_i$ and $-z = \sum_{i=k+1}^n x_i \otimes y_i$. Consequently, we have $\sum_{i=1}^n x_i \otimes y_i = 0$.

Now set $X := \operatorname{span}(x_1, \ldots, x_n) \subseteq E$ and $Y := \operatorname{span}(y_1, \ldots, y_n) \subseteq F$, and let $X_+ \subseteq X$ and $Y_+ \subseteq Y$ be the convex cones generated by $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$,

respectively. Note that X_+ is a closed proper cone in the finite-dimensional vector space X, since it is finitely generated (hence closed; see [AT07, Lemma 3.19]) and contained in the proper cone $X \cap E_+$ (hence also proper). Similarly, Y_+ is a closed proper cone in Y.

It follows from Corollary 9.9 that X_+^* and Y_+^* are generating cones in X^* and Y^*, respectively. Therefore clearly $X_+^* \otimes^\pi Y_+^*$ is generating in $X^* \otimes Y^*$. Since $\langle x \otimes y, \varphi \otimes \psi\rangle = \langle x, \varphi\rangle\langle y, \psi\rangle \geq 0$ for all $x \in X_+$, $y \in Y_+$, $\varphi \in X_+^*$, $\psi \in Y_+^*$, we have $X_+^* \otimes^\pi Y_+^* \subseteq (X_+ \otimes^\pi Y_+)^*$. It follows that $(X_+ \otimes^\pi Y_+)^*$ is also generating, and therefore $(X_+ \otimes^\pi Y_+)^{**} = \overline{X_+ \otimes^\pi Y_+}$ is a proper cone, by Lemma 9.8. Since $z, -z \in X_+ \otimes^\pi Y_+ \subseteq (X_+ \otimes^\pi Y_+)^{**}$, it follows that $z = 0$. ■

Remark 9.11. The final steps in the proof of Theorem 9.10 can be simplified with well-known results from the finite-dimensional theory, but we didn't need that. The dual of the projective cone $X_+ \otimes^\pi Y_+$ is the injective cone $X_+^* \otimes^\varepsilon Y_+^*$, and $X_+ \otimes^\pi Y_+$ is automatically closed, by [Tam77b] (see also Theorem 12.10 below).

Remark 9.12. In the proof of Theorem 9.10, we reduced the problem to finitely generated proper cones. There are many ways to prove this special case. Apart from the method used here and the proofs given in [DS70] and [Wor19], we could also have applied either one of the sufficient criteria from [PS69, Proposition 2.4]. Yet another method is mentioned in Remark 12.6.

Theorem 9.10 will be extended in Corollary 9.17 below, where we determine the lineality space of $E_+ \otimes^\pi F_+$ for arbitrary convex cones E_+, F_+. Furthermore, a result very similar to Theorem 9.10, giving criteria for $E_+ \otimes^\pi F_+$ to be semisimple (i.e. contained in a weakly closed proper cone), will be given in Corollary 11.11.

9.4 Faces of the projective cone

As a simple application of the theory developed so far, we develop two ways to combine faces of E_+ and F_+ to form a face of $E_+ \otimes^\pi F_+$. For closed, proper and generating cones in finite-dimensional spaces, one of these constructions was already pointed out (without proof) by Tam in [Tam77a, p. 53] and [Tam92, p. 71]. He likely had a different proof in mind which does not work in general; see Remark 9.20.

First we carry out the following very general construction; more convenient formulas and special cases will be studied afterwards.

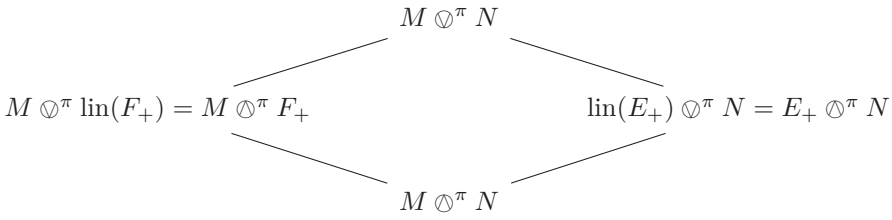
Theorem 9.13. *Let E, F be vector spaces, let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones, and let $M \subseteq E_+$, $N \subseteq F_+$ be non-empty faces. Define*

$$M \circledcirc^\pi N := (M \otimes^\pi F_+) + (E_+ \otimes^\pi N);$$
$$M \circledcirc^\pi N := (M \otimes^\pi N) + (\operatorname{lin}(E_+) \otimes^\pi F_+) + (E_+ \otimes^\pi \operatorname{lin}(F_+)).$$

Then:

(a) $M \circledcirc^\pi N$ and $M \circledcirc^\pi N$ are faces of $E_+ \otimes^\pi F_+$.

(b) The face lattice of $E_+ \otimes^\pi F_+$ contains the following sublattice:



Furthermore, $M \mathbin{⦶}^\pi N$ is not just the face generated by $M \mathbin{⦸}^\pi F_+$ and $E_+ \mathbin{⦸}^\pi N$, but even the sum of these faces, so we have

$$M \mathbin{⦶}^\pi N = (M \mathbin{⦶}^\pi \operatorname{lin}(F_+)) + (\operatorname{lin}(E_+) \mathbin{⦶}^\pi N) = (M \mathbin{⦸}^\pi F_+) + (E_+ \mathbin{⦸}^\pi N);$$
$$M \mathbin{⦸}^\pi N = (M \mathbin{⦶}^\pi \operatorname{lin}(F_+)) \cap (\operatorname{lin}(E_+) \mathbin{⦶}^\pi N) = (M \mathbin{⦸}^\pi F_+) \cap (E_+ \mathbin{⦸}^\pi N).$$

Assume furthermore that E, F, and $E \otimes F$ belong to the dual pairs $\langle E, E' \rangle$, $\langle F, F' \rangle$, and $\langle E \otimes F, G \rangle$, where G is a reasonable dual (i.e. $E' \otimes F' \subseteq G \subseteq \mathfrak{Bil}(E \times F)$). Then:

(c) If M and N are dual (resp. exposed) faces, then $M \mathbin{⦶}^\pi N$ is a dual (resp. exposed) face of $E_+ \otimes^\pi F_+$.

(d) If M and N as well as $\operatorname{lin}(E_+)$ and $\operatorname{lin}(F_+)$ are dual (resp. exposed) faces, then $M \mathbin{⦸}^\pi N$ is a dual (resp. exposed) face of $E_+ \otimes^\pi F_+$.

A mnemonic for the chosen notation: $M \mathbin{⦶}^\pi N$ is generated by the elements $x \otimes y \in E_+ \otimes_s F_+$ with $x \in M$ or $y \in N$, whereas $M \mathbin{⦸}^\pi N$ is generated by the elements $x \otimes y \in E_+ \otimes_s F_+$ with $x \in M$ and $y \in N$, together with what turns out to be the lineality space of $E_+ \otimes^\pi F_+$ (see Corollary 9.18 below).

Proof of Theorem 9.13.

(a) Let $I \subseteq E$ be an order ideal such that $M = I \cap E_+$ (e.g. $I = \operatorname{span}(M)$; see Proposition A.3(a)). Then the quotient cone $(E/I)_+ \subseteq E/I$ is proper, the natural map $\pi_I : E \to E/I$ is positive, and $M = \ker(\pi_I) \cap E_+$. Similarly, let $J \subseteq F$ be an ideal such that $N = J \cap F_+$; then $\pi_J : F \to F/J$ is a positive map to a space with a proper cone, and $N = \ker(\pi_J) \cap F_+$.

Now consider the linear map $\pi_I \otimes \pi_J : E \otimes F \to E/I \otimes F/J$. It follows from Proposition 9.2 that $\pi_I \otimes \pi_J$ is positive, and it follows from Theorem 9.10 that $(E/I)_+ \otimes^\pi (F/J)_+$ is a proper cone in $E/I \otimes F/J$, so $\ker(\pi_I \otimes \pi_J) \cap (E_+ \otimes^\pi F_+)$ is a face of $E_+ \otimes^\pi F_+$ (see Proposition A.4(b)). We claim that

$$\ker(\pi_I \otimes \pi_J) \cap (E_+ \otimes^\pi F_+) = M \mathbin{⦶}^\pi N. \tag{9.14}$$

Indeed, if $z = \sum_{i=1}^k x_i \otimes y_i$ with $x_1, \ldots, x_k \in E_+$, $y_1, \ldots, y_k \in F_+$ is such that $(\pi_I \otimes \pi_J)(z) = 0$, then we must have $(\pi_I \otimes \pi_J)(x_i \otimes y_i) = 0$ for all i (since $(E/I)_+ \otimes^\pi (F/J)_+$ is proper). As such, for each i we must have $x_i \in \ker(\pi_I) = I$ or $y_i \in \ker(\pi_J) = J$, or possibly both. Equivalently: $x_i \in I \cap E_+ = M$ or

$y_i \in J \cap F_+ = N$. This proves our claim (9.14), and we conclude that $M \mathbin{\textcircled{\scriptsize$\vee$}}^\pi N$ is a face of $E_+ \otimes^\pi F_+$.

To see that $M \mathbin{\textcircled{\scriptsize$\wedge$}}^\pi N$ is a face, we proceed analogously, where the linear map $\pi_I \otimes \pi_J$ is replaced by the linear map

$$\begin{aligned} Q_{I,J} : E \otimes F &\to (E/I \otimes F/\operatorname{lin}(F_+)) \oplus (E/\operatorname{lin}(E_+) \otimes F/J), \\ x \otimes y &\mapsto (\pi_I(x) \otimes \pi_{\operatorname{lin}(F_+)}(y)) \oplus (\pi_{\operatorname{lin}(E_+)}(x) \otimes \pi_J(y)). \end{aligned}$$

If $z = \sum_{i=1}^k x_i \otimes y_i$ with $x_1, \dots, x_k \in E_+$, $y_1, \dots, y_k \in F_+$ and $Q_{I,J}(z) = 0$, then again we must have $Q_{I,J}(x_i \otimes y_i) = 0$ for all i (since $Q_{I,J}$ is positive and the cone in the codomain is proper). Then either $x_i \in \ell(E_+) \subseteq M$, or $y_i \in \ell(F_+) \subseteq N$, or $x_i \notin \ell(E_+)$ and $y_i \notin \ell(F_+)$. In the latter case, we must have $x_i \in M$ and $y_i \in N$. This way we find

$$\ker(Q_{I,J}) \cap (E_+ \otimes^\pi F_+) = M \mathbin{\textcircled{\scriptsize$\wedge$}}^\pi N.$$

It follows that $M \mathbin{\textcircled{\scriptsize$\wedge$}}^\pi N$ is also a face of $E_+ \otimes^\pi F_+$.

(b) Using the notation from the proof of (a), note that

$$\ker(Q_{I,J}) = \ker(\pi_I \otimes \pi_{\operatorname{lin}(F_+)}) \cap \ker(\pi_{\operatorname{lin}(E_+)} \otimes \pi_J).$$

It follows that

$$M \mathbin{\textcircled{\scriptsize$\wedge$}}^\pi N = (M \mathbin{\textcircled{\scriptsize$\vee$}}^\pi \operatorname{lin}(F_+)) \cap (\operatorname{lin}(E_+) \mathbin{\textcircled{\scriptsize$\vee$}}^\pi N).$$

The other formulas follow straight from the definitions: we have

$$\begin{aligned} (M \mathbin{\textcircled{\scriptsize$\vee$}}^\pi \operatorname{lin}(F_+)) + (\operatorname{lin}(E_+) \mathbin{\textcircled{\scriptsize$\vee$}}^\pi N) &= (M \otimes^\pi F_+) + (E_+ \otimes^\pi \operatorname{lin}(F_+)) \\ &\quad + (\operatorname{lin}(E_+) \otimes^\pi F_+) + (E_+ \otimes^\pi N) \\ &= (M \otimes^\pi F_+) + (E_+ \otimes^\pi N) \\ &= M \mathbin{\textcircled{\scriptsize$\vee$}}^\pi N, \end{aligned}$$

since $\operatorname{lin}(E_+) \subseteq M$ and $\operatorname{lin}(F_+) \subseteq N$. Likewise,

$$\begin{aligned} M \mathbin{\textcircled{\scriptsize$\wedge$}}^\pi F_+ &= (M \otimes^\pi F_+) + (\operatorname{lin}(E_+) \otimes^\pi F_+) + (E_+ \otimes^\pi \operatorname{lin}(F_+)) \\ &= (M \otimes^\pi F_+) + (E_+ \otimes^\pi \operatorname{lin}(F_+)) \\ &= M \mathbin{\textcircled{\scriptsize$\vee$}}^\pi \operatorname{lin}(F_+), \end{aligned}$$

and the formula $E_+ \mathbin{\textcircled{\scriptsize$\wedge$}}^\pi N = \operatorname{lin}(E_+) \mathbin{\textcircled{\scriptsize$\vee$}}^\pi N$ follows analogously.

(c) If $M = {}^\diamond M_1$ and $N = {}^\diamond N_1$, then it is routinely verified that $M \mathbin{\textcircled{\scriptsize$\vee$}}^\pi N = {}^\diamond(M_1 \otimes_s N_1)$. If M and N are furthermore exposed, then we may take M_1 and N_1 to be singletons; consequently, $M_1 \otimes_s N_1$ is also a singleton.

(d) This follows from (c) and the intersection formula from (b). ■

Remark 9.15. In Theorem 9.13(d), it is required that $\lin(E_+)$ and $\lin(F_+)$ are exposed/dual faces. Sometimes this is automatically the case. If E_+ is weakly closed, then $\lin(E_+) = \lin(\overline{E_+}^w) = {}^\perp(E'_+) = {}^\circ(E'_+)$, so in this case $\lin(E_+)$ is always a dual face. Likewise, if E is a separable normed space and E_+ is closed, then $\lin(E_+)$ is automatically exposed; see Corollary A.19.

To see that this assumption cannot be omitted, let $E := \mathbb{R}^2$ with the lexicographical cone, and let $F := \mathbb{R}$ with the standard cone. Then the unique one-dimensional face $M \subseteq E_+$ and the trivial face $N := \{0\} \subseteq \mathbb{R}$ are both exposed (hence dual), but $M \mathbin{\triangle^\pi} N = \{0\}$ is neither exposed nor dual in $E_+ \otimes^\pi F_+ \cong E_+$.

Remark 9.16. By dualizing the example from Example 10.51 below, one can show that not every facet of $E_+ \otimes^\pi F_+$ is necessarily of the form $M \mathbin{\triangledown^\pi} N$ or $M \mathbin{\triangle^\pi} N$. In follows that, in general, not every face of $E_+ \otimes^\pi F_+$ can be written as an intersection of faces of the type $M \mathbin{\triangledown^\pi} N$ or $M \mathbin{\triangle^\pi} N$.

We proceed to point out the consequences of Theorem 9.13. First of all, it allows us to extend Theorem 9.10, giving a direct formula for the lineality space of $E_+ \otimes^\pi F_+$.

Corollary 9.17 (The lineality space of the projective cone)**.** *Let E and F be vector spaces, and let $E_+ \subseteq E$ and $F_+ \subseteq F$ be convex cones. Then one has*

$$\begin{aligned}\lin(E_+ \otimes^\pi F_+) &= (\lin(E_+) \otimes^\pi F_+) + (E_+ \otimes^\pi \lin(F_+)) \\ &= (\lin(E_+) \otimes \operatorname{span}(F_+)) + (\operatorname{span}(E_+) \otimes \lin(F_+)).\end{aligned}$$

Proof. If $x \in \lin(E_+)$ and $y \in F_+$, then $\pm x \otimes y \in E_+ \otimes^\pi F_+$, so $x \otimes y \in \lin(E_+ \otimes^\pi F_+)$. Similarly, if $x \in E_+$ and $y \in \lin(F_+)$, then $x \otimes y \in \lin(E_+ \otimes^\pi F_+)$, so we have

$$(\lin(E_+) \otimes^\pi F_+) + (E_+ \otimes^\pi \lin(F_+)) \subseteq \lin(E_+ \otimes^\pi F_+).$$

Conversely, it follows from Theorem 9.13(a) that the 'upper face' $\lin(E_+) \mathbin{\triangledown^\pi} \lin(F_+) = (\lin(E_+) \otimes^\pi F_+) + (E_+ \otimes^\pi \lin(F_+))$ is a face of $E_+ \otimes^\pi F_+$, so it must contain the minimal face $\lin(E_+ \otimes^\pi F_+)$. The first equality follows.

For the second equality, we claim that $\lin(E_+) \otimes^\pi F_+ = \lin(E_+) \otimes \operatorname{span}(F_+)$. Indeed, for $x \in \lin(E_+)$ and $y \in \operatorname{span}(F_+)$ we may write $y = u - v$ (for some $u, v \in F_+$), so we have $x \otimes y = (x \otimes u) + ((-x) \otimes v) \in E_+ \otimes^\pi F_+$. Taking positive linear combinations proves our claim. Analogously, we have $E_+ \otimes^\pi \lin(F_+) = \operatorname{span}(E_+) \otimes \lin(F_+)$, and the second equality follows. ∎

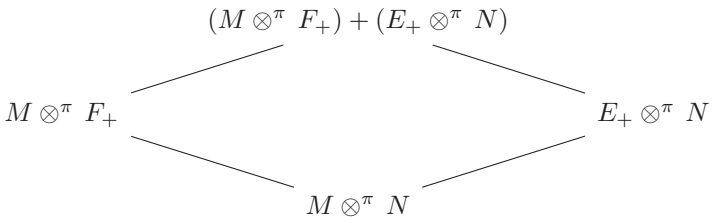
This direct formula for the lineality space also simplifies the formula for the lower face $M \mathbin{\triangle^\pi} N$.

Corollary 9.18. *Let E, F be vector spaces, let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones, and let $M \subseteq E_+$, $N \subseteq F_+$ be non-empty faces. Then one has*

$$M \mathbin{\triangle^\pi} N = (M \otimes^\pi N) + \lin(E_+ \otimes^\pi F_+),$$

and this defines a face of $E_+ \otimes^\pi F_+$.

In particular, if E_+ and F_+ are proper cones, then $M \otimes^\pi N$ is a face of $E_+ \otimes^\pi F_+$, and the sublattice from Theorem 9.13(b) reduces to



For closed, proper and generating cones in finite-dimensional spaces, the fact that $M \otimes^\pi N$ is a face of $E_+ \otimes^\pi F_+$ was already pointed out (without proof) by Tam in [Tam77a, p. 53] and [Tam92, p. 71]. He likely had a different proof in mind, which we outline in Remark 9.20 below.

Remark 9.19. In general, $M \otimes^\pi N$ is not a face of $E_+ \otimes^\pi F_+$. If E_+ or F_+ is not proper, then the term $+\lin(E_+ \otimes^\pi F_+)$ cannot be omitted in Corollary 9.18. Indeed, suppose that E_+ is not proper. Choose $x \in \lin(E_+) \setminus \{0\}$ and $y \in \operatorname{span}(F_+) \setminus \operatorname{span}(N)$. Then $x \otimes y \in \lin(E_+ \otimes^\pi F_+)$, by Corollary 9.17. However, $x \otimes y \notin M \otimes^\pi N$, so $M \otimes^\pi N$ is not a face, because every face must contain the lineality space.

Remark 9.20. If E_+^* and F_+^* separate points on E and F,[1] then there is a simpler way to show that $M \otimes^\pi N$ is a face of $E_+ \otimes^\pi F_+$. Indeed, let $z, z' \in E_+ \otimes^\pi F_+$ be such that $z'' := z + z' \in M \otimes^\pi N$, and write $z = \sum_{i=1}^k x_i \otimes y_i$, where $x_1, \dots, x_k \in E_+$ and $y_1, \dots, y_k \in F_+$ are all non-zero. For $i \in \{1, \dots, k\}$, choose $\varphi_i \in E_+^*$ and $\psi_i \in F_+^*$ such that $\varphi_i(x_i), \psi_i(y_i) > 0$. Then we have $0 < \varphi_i(x_i) y_i \leq \sum_{j=1}^k \varphi_i(x_j) y_j = (\varphi_i \otimes \mathrm{id}_F)(z) \leq (\varphi_i \otimes \mathrm{id}_F)(z'') \in N$, hence $y_i \in N$. Likewise, $0 < \psi_i(y_i) x_i \leq (\mathrm{id}_E \otimes \psi_i)(z'') \in M$, hence $x_i \in M$. It follows that $z \in M \otimes^\pi N$, which shows that $M \otimes^\pi N$ is a face.

In particular, this simple proof settles the case when E and F are finite-dimensional and E_+ and F_+ are closed, proper and generating. This special case was already pointed out (without proof) by Tam in [Tam77a, p. 53] and [Tam92, p. 71]. The proof he had in mind is probably similar to short proof given here.

As a final application, we note that Theorem 9.13 is also a statement about preservation of bipositive maps.

Proposition 9.21. *Let E and F be vector spaces, and let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones. If E_+ and F_+ are proper and if $I \subseteq E$, $J \subseteq F$ are ideals, then the inclusion $I \otimes J \hookrightarrow E \otimes F$ is bipositive (with respect to the projective cone).*

Proof. Let $Q_{I,J} : E \otimes F \to (E/I \otimes F) \oplus (E \otimes F/J)$ be the map from the proof of Theorem 9.13(a). It follows from said proof (and Corollary 9.18) that $I_+ \otimes^\pi J_+ = \ker(Q_{I,J}) \cap (E_+ \otimes^\pi F_+)$. To complete the proof, note that $\ker(Q_{I,J}) = I \otimes J$. ∎

Example 9.6 shows that this is not true if one of the cones is not proper.

[1] In other words, E_+ and F_+ are semisimple with respect to the dual pairs $\langle E, E^* \rangle$ and $\langle F, F^* \rangle$; see [Dob20a]. Schaefer [Sch58] called such cones *regular*.

9.5 Extremal rays of the projective cone

The results from §9.4 show us how to construct faces in the projective tensor cone, even though not all faces are reached this way (see Remark 9.16). Nevertheless, it turns out that all extremal rays of $E_+ \otimes^\pi F_+$ are obtained in this way.

Recall that $\operatorname{rext}(E_+) \subseteq E_+ \setminus \{0\}$ denotes the set of extremal directions, and $M \otimes_s N$ denotes the entry-wise tensor product $\{x \otimes y : x \in M,\ y \in N\}$.

Theorem 9.22 (The extremal rays of the projective cone)**.** *Let E, F be vector spaces equipped with convex cones $E_+ \subseteq E$, $F_+ \subseteq F$. Then*

$$\operatorname{rext}(E_+ \otimes^\pi F_+) = \operatorname{rext}(E_+) \otimes_s \operatorname{rext}(F_+).$$

Proof. "$\subseteq$". Suppose that $z \in (E_+ \otimes^\pi F_+) \setminus \{0\}$ defines an extremal ray. Write $z = \sum_{i=1}^k x_i \otimes y_i$ with $x_1, \ldots, x_k \in E_+$, $y_1, \ldots, y_k \in F_+$, and $x_i \otimes y_i \neq 0$ for all $i \in [k]$. By extremality of z there are $\lambda_1, \ldots, \lambda_k \in \mathbb{R}_{>0}$ such that $\lambda_i x_i \otimes y_i = z$ $(i \in [k])$. In particular, $z = \lambda_1 x_1 \otimes y_1$. Now suppose that $0 \leq v \leq x_1$, then $0 \leq \lambda_1 v \otimes y_1 \leq z$, so by extremality of z we must have $\mu\lambda_1 v \otimes y_1 = z$ for some $\mu \in \mathbb{R}_{\geq 0}$. Since $y_1 \neq 0$ and $\lambda_1 \neq 0$, it follows that $\mu v = x_1$, so we see that x_1 defines an extremal ray of E_+. Analogously, y_1 defines an extremal ray of F_+. This proves the inclusion $\operatorname{rext}(E_+ \otimes^\pi F_+) \subseteq \operatorname{rext}(E_+) \otimes_s \operatorname{rext}(F_+)$.

"$\supseteq$". Let $x_0 \in E_+ \setminus \{0\}$ and $y_0 \in F_+ \setminus \{0\}$ define extremal rays in E_+ and F_+, respectively. Then $M := \{\lambda x_0 : \lambda \geq 0\}$ defines a face of E_+. Every face contains the lineality space, but M does not contain a non-zero subspace, so it follows that E_+ is a proper cone. Analogously, $N := \{\mu y_0 : \mu \geq 0\}$ defines a face of F_+, so F_+ is proper. Now it follows from Corollary 9.18 that $M \otimes^\pi N$ is a face of $E_+ \otimes^\pi F_+$. In other words: $x_0 \otimes y_0$ defines an extremal ray of $E_+ \otimes^\pi F_+$. ■

Remark 9.23. Remarkably, Theorem 9.22 has no corner cases: it is true for every pair of convex cones. In particular, if $\operatorname{rext}(E_+)$ or $\operatorname{rext}(F_+)$ is empty, then $\operatorname{rext}(E_+ \otimes^\pi F_+)$ is empty as well. Conversely, if each of E_+ and F_+ has an extremal ray, then so does $E_+ \otimes^\pi F_+$.[2]

Again, in the case where E and F are finite-dimensional and E_+ and F_+ are closed, proper and generating, this was already pointed out (without proof) by Tam in [Tam77a, p. 53] and [Tam92, p. 71]. See also Remark 9.20.

9.6 An application to tensor products of absolutely convex sets

We conclude our study of the projective cone with an application in convex geometry. Using a slight modification of the construction from §9.4, we show that faces of

[2] It should be noted that many standard cones in infinite-dimensional spaces do not have sufficiently many extremal rays to generate the cone. For instance, the positive cone of $C[0,1]$ has no extremal rays at all.

absolutely convex sets M and N determine faces of their tensor product $M \otimes^\pi N := \operatorname{conv}\{x \otimes y : x \in M,\ y \in N\}$.[3]

This application is based on the following general principle, giving sufficient conditions for the sum of faces $M_1 \oslash^\pi N_1$ and $M_2 \oslash^\pi N_2$ (see §9.4) to be another face in the projective cone $E_+ \otimes^\pi F_+$. (This is a vast generalization of the method of [BCG13, Example 3.7].)

Proposition 9.24. *Let E, F be vector spaces, let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones, and let $M_1, M_2 \subseteq E_+$ and $N_1, N_2 \subseteq F_+$ be faces. If $M_1 \cap M_2 = \lin(E_+)$ and $N_1 \cap N_2 = \lin(F_+)$, then*

$$(M_1 \oslash^\pi N_1) + (M_2 \oslash^\pi N_2) \;=\; (M_1 \otimes^\pi N_2) \cap (M_2 \otimes^\pi N_1).$$

In particular, in this case $(M_1 \oslash^\pi N_1) + (M_2 \oslash^\pi N_2)$ is a face of $E_+ \otimes^\pi F_+$.

Proof. "$\subseteq$". By Theorem 9.13(b), we have $M_1 \oslash^\pi N_1 \subseteq M_1 \oslash^\pi F_+ = M_1 \otimes^\pi \lin(F_+) \subseteq M_1 \otimes^\pi N_2$. Three analogous inclusions prove the forward inclusion.

"$\supseteq$". Let $z \in (M_1 \otimes^\pi N_2) \cap (M_2 \otimes^\pi N_1)$, and write $z = \sum_{i=1}^k x_i \otimes y_i$ with $x_1, \ldots, x_k \in E_+$ and $y_1, \ldots, y_k \in F_+$. Since $z \in M_1 \otimes^\pi N_2$, it follows from the proof of Theorem 9.13(a) that for all i we have $x_i \in M_1$ or $y_i \in N_2$, or possibly both. Likewise, for all i we have $x_i \in M_2$ or $y_i \in N_1$, or possibly both.

If $x_i \in \lin(E_+)$ or $y_i \in \lin(F_+)$, then $x_i \otimes y_i \in \lin(E_+ \otimes^\pi F_+) \subseteq (M_1 \oslash^\pi N_1) \cap (M_2 \oslash^\pi N_2)$, since every face contains the lineality space. So assume $x_i \notin \lin(E_+)$ and $y_i \notin \lin(F_+)$. Then, by assumption, x_i (resp. y_i) is contained in at most one of M_1 and M_2 (resp. N_1 and N_2). Combined with earlier constraints, this show that we must either have $x_i \in M_1 \setminus M_2$ and $y_i \in N_1 \setminus N_2$, or otherwise $x_i \in M_2 \setminus M_1$ and $y_i \in N_2 \setminus N_1$. Either way, $x_i \otimes y_i \in (M_1 \oslash^\pi N_1) + (M_2 \oslash^\pi N_2)$. ∎

If E is a vector space and $C \subseteq E$ is a convex subset, then the *homogenization* $\mathscr{C}(C)$ of C is the convex cone generated by $C \oplus \{1\} \subseteq E \oplus \mathbb{R}$. Note that $\mathscr{C}(C)$ is always a proper cone, and that the faces of C are in bijective correspondence with the faces of $\mathscr{C}(C)$.

Since we are working over the real numbers, a convex set $C \subseteq E$ is absolutely convex if and only if $C = -C$. For sets of this kind, there is a simple way to identify the projective tensor product of the homogenizations $\mathscr{C}(C)$ and $\mathscr{C}(D)$ with the homogenization of $\operatorname{conv}(C \otimes_s D)$:

Proposition 9.25. *Let E and F be (real) vector spaces and let $C \subseteq E$, $D \subseteq F$ be absolutely convex sets. Under the natural isomorphism $(E \oplus \mathbb{R}) \otimes (F \oplus \mathbb{R}) = (E \otimes F) \oplus E \oplus F \oplus \mathbb{R}$, one has*

$$(\mathscr{C}(C) \otimes^\pi \mathscr{C}(D)) \cap \big((E \otimes F) \oplus \{0\} \oplus \{0\} \oplus \{1\}\big) \;=\; \{(z, 0, 0, 1) : z \in \operatorname{conv}(C \otimes_s D)\}.$$

[3] Some authors define the projective tensor product of convex sets to be the *closed* convex hull of $M \otimes_s N$ (e.g. [AS17, §4.1.4]), but our methods are not equipped to deal with closures. See also Remark 9.30.

Proof. Under the aforementioned natural isomorphism, we have $(x,\lambda)\otimes(y,\mu) = (x\otimes y, \mu x, \lambda y, \lambda\mu)$.

"$\subseteq$". Let $(z,0,0,1)\in \mathscr{C}(C)\otimes^\pi \mathscr{C}(D)$ be given, and write $(z,0,0,1) = \sum_{i=1}^k \lambda_i \cdot (x_i,1)\otimes(y_i,1)$ with $\lambda_1,\dots,\lambda_k \geq 0$, $x_1,\dots,x_k \in C$ and $y_1,\dots,y_k\in D$. Then $(z,0,0,1) = \sum_{i=1}^k \lambda_i\cdot(x_i\otimes y_i, x_i, y_i, 1)$, so we have $\sum_{i=1}^k\lambda_i = 1$ and $z = \sum_{i=1}^k \lambda_i x_i \otimes y_i \in \operatorname{conv}(C\otimes_s D)$.

"$\supseteq$". Let $z\in\operatorname{conv}(C\otimes_s D)$ be given, and write $z = \sum_{i=1}^k \lambda_i x_i\otimes y_i$ with $x_1,\dots,x_k\in C$, $y_1,\dots,y_k\in D$, $\lambda_1,\dots,\lambda_k\geq 0$, and $\sum_{i=1}^k\lambda_i = 1$. Since $(x_i,1)\otimes(y_i,1) + (-x_i,1)\otimes(-y_i,1) = 2(x_i\otimes y_i,0,0,1)$, we may write

$$(z,0,0,1) = \sum_{i=1}^k \tfrac{1}{2}\lambda_i\cdot\big((x_i,1)\otimes(y_i,1) + (-x_i,1)\otimes(-y_i,1)\big). \tag{9.26}$$

Since C and D are absolutely convex, we have $(\pm x_i,1)\in\mathscr{C}(C)$ and $(\pm y_i,1)\in\mathscr{C}(D)$ for all $i\in\{1,\dots,k\}$, hence $(z,0,0,1)\in\mathscr{C}(C)\otimes^\pi\mathscr{C}(D)$. ∎

Theorem 9.27. *Let E and F be (real) vector spaces, let $C\subseteq E$, $D\subseteq F$ be absolutely convex, and let $M\subset C$, $N\subset D$ be proper faces. Then* $\operatorname{conv}(M\otimes_s N)$ *is a face of* $\operatorname{conv}(C\otimes_s D)$.

Proof. By symmetry, $-M\subseteq C$ and $-N\subseteq D$ also define faces of C and D. First we prove that $M\cap -M = \varnothing$. Suppose that $x\in M\cap -M$. Then also $-x\in M\cap -M$, so by convexity $0\in M\cap -M$. But then for every $y\in C$ we must have $y,-y\in M$, since 0 belongs to the relative interior of the line segment joining y and $-y$. This contradicts our assumption that M is a proper face, so we conclude that $M\cap -M=\varnothing$. Analogously, $N\cap -N=\varnothing$.

Let $M_1\subseteq\mathscr{C}(C)$ be the face of $\mathscr{C}(C)$ associated with M, and let $M_2\subseteq\mathscr{C}(C)$ be the face associated with $-M$. Since $M\cap -M=\varnothing$, it follows that $M_1\cap M_2 = \{0\}$. Similarly, let N_1 and N_2 be the faces of $\mathscr{C}(D)$ associated with N and $-N$, respectively; then $N_1\cap N_2=\{0\}$. Hence it follows from Proposition 9.24 that $(M_1\otimes^\pi N_1)+(M_2\otimes^\pi N_2)$ is a face of $\mathscr{C}(C)\otimes^\pi\mathscr{C}(D)$. To complete the proof, we show that

$$\big((M_1\otimes^\pi N_1)+(M_2\otimes^\pi N_2)\big)\cap\big((E\otimes F)\oplus\{0\}\oplus\{0\}\oplus\{1\}\big) = \{(z,0,0,1) : z\in\operatorname{conv}(M\otimes_s N)\}.$$

We proceed analogously to the proof of Proposition 9.25.

"$\subseteq$". Let $(z,0,0,1)\in(M_1\otimes^\pi N_1)+(M_2\otimes^\pi N_2)$ be given. Then we may choose integers $n\geq k\geq 0$, scalars $\lambda_1,\dots,\lambda_n\geq 0$ and vectors $x_1,\dots,x_n\in M$, $y_1,\dots,y_n\in N$ such that $(z,0,0,1) = \sum_{i=1}^k\lambda_i\cdot(x_i,1)\otimes(y_i,1) + \sum_{i=k+1}^n\lambda_i\cdot(-x_i,1)\otimes(-y_i,1)$. Therefore $\sum_{i=1}^n\lambda_i = 1$ and $z=\sum_{i=1}^n\lambda_i x_i\otimes y_i$, which shows that $z\in\operatorname{conv}(M\otimes_s N)$.

"$\supseteq$". Let $z\in\operatorname{conv}(M\otimes_s N)$ be given, and write $z=\sum_{i=1}^k\lambda_i x_i\otimes y_i$ with $x_1,\dots,x_k\in M$, $y_1,\dots,y_k\in N$, $\lambda_1,\dots,\lambda_k\geq 0$, and $\sum_{i=1}^k\lambda_i = 1$. Then it follows from (9.26) that $(z,0,0,1)\in(M_1\otimes^\pi N_1)+(M_2\otimes^\pi N_2)$. ∎

Corollary 9.28. *Let E and F be (real) vector spaces, let $C\subseteq E$, $D\subseteq F$ be absolutely convex, and let $x_0\in C$, $y_0\in D$ be extreme points. Then $x_0\otimes y_0$ is an extreme point of* $\operatorname{conv}(C\otimes_s D)$.

Remark 9.29. Theorem 9.27 fails if one of the faces is not proper. Indeed, if $M = C$, then $0 \in M \otimes_s N$, so now $\operatorname{conv}(M \otimes_s N)$ is a face only if $\operatorname{conv}(M \otimes_s N) = \operatorname{conv}(C \otimes_s D)$.

Furthermore, Theorem 9.27 and Corollary 9.28 do not hold for non-symmetric convex sets. For example, $1 \otimes 2$ is not an extreme point of $\operatorname{conv}([-1,1] \otimes_s [2,3]) \subseteq \mathbb{R} \otimes \mathbb{R} = \mathbb{R}$.

Remark 9.30. In many applications it is natural to start with *closed* absolutely convex sets, and take the *closed* convex hull of their tensor product (e.g. [PTT11, Remark 3.19], [AS17, §4.1.4], or when computing the closed unit ball of the projective norm). Our methods are not equipped to deal with closures.

If E, F are finite-dimensional and if C, D are compact, then $\operatorname{conv}(C \otimes_s D)$ is automatically compact, so here taking closures is not necessary. In particular:

Corollary 9.31. *Let E and F be (real) finite-dimensional normed spaces. Then the closed unit ball of the projective norm preserves proper faces: if $M \subset B_E$, $N \subset B_F$ are proper faces, then* $\operatorname{conv}(M \otimes_s N)$ *is a face of $B_{E \otimes_\pi F}$.*

This had already been known for extreme points. More generally, if E and F are Banach spaces, then it follows from a result of Tseitlin [Tse76] (see also [RS82]) that the closed unit ball of the *completed* projective tensor product $E' \mathbin{\tilde{\otimes}_\pi} F'$ preserves extreme points, provided that E' or F' has the approximation property and E' or F' has the Radon–Nikodym property.[4] In particular, this settles the finite-dimensional case, proving Corollary 9.31 for extreme points.

Remark 9.32. We do not know whether the closed unit ball of the projective norm always preserves extreme points, even in the algebraic tensor product. This does not follow from Corollary 9.28, because the closed unit ball of $E \otimes_\pi F$ is the *closure* of $\operatorname{conv}(B_E \otimes_s B_F)$. Known results in this direction usually start with something stronger than an extreme point, such as a *denting point* (see [RS86b, Theorem 5], [Wer87, Corollary 4]).

We suspect that there are Banach spaces E and F such that the projective norm does not preserve all extreme points of their closed unit balls, but we have not been able to construct such an example. To our knowledge, no such examples are known in the literature either.

Finally, we should point out that the injective norm does not preserve extreme points; see Remark 10.52.

[4] The cited results relate to extreme points in duals of operator spaces. Our assumptions on E' and F' ensure that $E' \mathbin{\tilde{\otimes}_\pi} F' \cong (E \mathbin{\tilde{\otimes}_\varepsilon} F)'$ isometrically; see [DF93, Theorem 16.6].

Chapter 10

The injective cone

In this chapter, we carry out an in-depth study of the properties of the injective cone. This cone depends not only on the vector spaces E, F and the cones E_+, F_+, but also on the dual spaces E', F'. Therefore we will work with dual pairs.

This chapter is based on Chapter 4 of [Dob20b].

Introduction

Let $\langle E, E' \rangle$, $\langle F, F' \rangle$ be dual pairs of (real) vector spaces, and let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones in the primal spaces. The *injective cone*[1] in $E \otimes F$ is defined as

$$E_+ \otimes^\varepsilon F_+ := \big\{ u \in E \otimes F \,:\, \langle u, \varphi \otimes \psi \rangle \geq 0 \text{ for all } \varphi \in E'_+,\, \psi \in F'_+ \big\}.$$

The notation causes some ambiguity, because $E_+ \otimes^\varepsilon F_+$ does not only depend on E_+ and F_+, but also on the dual pairs $\langle E, E' \rangle$ and $\langle F, F' \rangle$. To be fully precise, the injective cone should be denoted as something like $(\langle E, E' \rangle, E_+) \otimes^\varepsilon (\langle F, F' \rangle, F_+)$. We forego this cumbersome notation for the sake of clarity; it will always be clear what is meant.

If E and F are locally convex and if $E \otimes F$ is equipped with a compatible topology α (in the sense of Grothendieck [Gro55, p. 89], see also [Köt79, §44.1]), then for every $\varphi \in E'$, $\psi \in F'$ the tensor product $\varphi \otimes \psi : E \otimes_\alpha F \to \mathbb{R}$ is continuous, and as such has a unique extension to $E \,\tilde{\otimes}_\alpha\, F$. In this setting we may likewise define the injective cone as

$$E_+ \,\tilde{\otimes}^\varepsilon_\alpha\, F_+ := \big\{ u \in E \,\tilde{\otimes}_\alpha\, F \,:\, (\varphi \,\tilde{\otimes}_\alpha\, \psi)(u) \geq 0 \text{ for all } \varphi \in E'_+,\, \psi \in F'_+ \big\}.$$

Clearly $E_+ \otimes^\varepsilon F_+ = (E_+ \,\tilde{\otimes}^\varepsilon_\alpha\, F_+) \cap (E \otimes F)$. Note that, unlike the projective cone, the injective cone typically becomes larger when passing from the algebraic tensor product $E \otimes F$ to the completion $E \,\tilde{\otimes}_\alpha\, F$.

[1] A note about terminology: in the literature, $E_+ \otimes^\varepsilon F_+$ is usually called the *biprojective cone* (see e.g. [Mer64, PS69, Bir76]). The results in this section show that this cone is in many ways analogous to the injective topology, and as such deserves the name *injective cone*. The only prior use of this name (that we are aware of) is in [Wit74] and [Mul97].

Remark 10.1. Let G be any reasonable dual of $E \otimes F$ (cf. page 109). It is clear from the definition that $E_+ \otimes^\varepsilon F_+$ is the predual cone of $E'_+ \otimes^\pi F'_+$ under the dual pairing $\langle E \otimes F, G\rangle$. Likewise, $E_+ \tilde{\otimes}^\varepsilon_\alpha F_+ \subseteq E \tilde{\otimes}_\alpha F$ is the predual cone of $E'_+ \otimes^\pi F'_+ \subseteq (E \tilde{\otimes}_\alpha F)'$.

An immediate consequence is that the injective cone is always weakly closed. Furthermore, by the bipolar theorem, the dual cone of $E_+ \otimes^\varepsilon F_+$ with respect to the dual pair $\langle E \otimes F, G\rangle$ is the $\sigma(G, E \otimes F)$-closure of $E'_+ \otimes^\pi F'_+$. (Note that this need not be contained in $E' \otimes F'$.) Similarly, in the locally convex setting, the dual cone of $E_+ \tilde{\otimes}^\varepsilon_\alpha F_+$ is the weak-$*$ closure of $E'_+ \otimes^\pi F'_+ \subseteq (E \tilde{\otimes}_\alpha F)'$.

In this chapter, we give a detailed study of the properties of the injective cone. In §10.1, we establish the characteristic property of the injective cone. In §10.2, we show that the injective cone preserves positive maps, bipositive maps, and retracts, but not pushforwards. In §10.3 we determine necessary and sufficient conditions for the injective cone to be proper. Finally, in §10.4–§10.6 we show how faces in E_+ and F_+ determine faces of $E_+ \otimes^\varepsilon F_+$.

10.1 The characteristic property of the injective cone

We show that the injective cone can be identified with a cone of positive bilinear forms. Let $E \circledast F$ denote the space of separately weak-$*$ continuous bilinear forms on $E' \times F'$:

$$E \circledast F := \mathfrak{Bil}\left(E'_{w*} \times F'_{w*}\right).$$

(Köthe [Köt79, §44.4] uses the symbol $\boxtimes$ instead of $\circledast$.)

We shall understand $E \circledast F$ to be equipped with the cone it inherits from $\mathrm{Bil}(E' \times F')$. In other words, $b \in E \circledast F$ is positive if and only if $b(\varphi, \psi) \geq 0$ for all $\varphi \in E'_+$, $\psi \in F'_+$.

The characteristic property of the injective cone is that it is given by a bipositive map to $E \circledast F$ (algebraic case) or $\tilde{E} \circledast \tilde{F}$ (completed locally convex case).

Remark 10.2. Statements about positive bilinear forms can be turned into equivalent statements about positive linear operators in the following way. Recall that $\mathfrak{Bil}(E'_{w*} \times F'_{w*})$ is naturally isomorphic to $\mathfrak{L}(E'_{w*}, F_w)$. Under this correspondence, the positive cone of $\mathfrak{Bil}(E'_{w*} \times F'_{w*})$ is the cone of *approximately positive* operators $E'_{w*} \to F_w$, i.e. those operators T that satisfy $T[E'_+] \subseteq \overline{F_+}^w$. In particular, if F_+ is weakly closed, then this is just the cone of positive operators $E'_{w*} \to F_w$. Similarly, $\mathfrak{Bil}(E'_{w*} \times F'_{w*}) \cong \mathfrak{L}(F'_{w*}, E_w)$, and the positive cone of $\mathfrak{Bil}(E'_{w*} \times F'_{w*})$ corresponds with the approximately positive cone of $\mathfrak{L}(F'_{w*}, E_w)$.

The advantage of sticking to bilinear forms is twofold: it keeps the theory symmetric in E and F, and it avoids the nuisance of having to take the weak closure of F_+ (or E_+).

We proceed to prove the characteristic property in three settings: the algebraic tensor product, the completed injective tensor product, and arbitrary completed tensor products.

Situation I: the algebraic tensor product

Let $\langle E, E'\rangle$ and $\langle F, F'\rangle$ be dual pairs. Equip E' and F' with their respective weak-$*$ topologies, and denote these spaces as E'_{w*} and F'_{w*}. The dual pairing $\langle E\otimes F, E'\otimes F'\rangle$ yields a natural map $E\otimes F \hookrightarrow (E'\otimes F')^* \cong \mathrm{Bil}(E'\times F')$. Note that the elements of $E\otimes F$ give rise to jointly continuous bilinear maps $E'_{w*}\times F'_{w*}\to\mathbb{R}$. Indeed, an elementary tensor $x_0\otimes y_0\in E\otimes F$ defines the bilinear map $(\varphi,\psi)\mapsto\langle x_0,\varphi\rangle\langle y_0,\psi\rangle$, which is easily seen to be jointly continuous (use that $\varphi\mapsto\langle x_0,\varphi\rangle$ and $\psi\mapsto\langle y_0,\psi\rangle$ are continuous). Consequently, finite sums of elementary tensors also define jointly continuous bilinear maps, and the claim follows. This gives us natural inclusions

$$E\otimes F\subseteq \mathscr{B}\!il(E'_{w*}\times F'_{w*})\subseteq E\circledast F\subseteq \mathrm{Bil}(E'\times F'). \tag{10.3}$$

From left to right, these are the spaces of (continuous) finite rank, jointly continuous, separately continuous, and all bilinear forms on $E'_{w*}\times F'_{w*}$.

Proposition 10.4. *The elements of $E_+\otimes^\varepsilon F_+$ are precisely those elements in $E\otimes F$ which define a positive bilinear map $E'\times F'\to\mathbb{R}$; that is:*

$$E_+\otimes^\varepsilon F_+ = \mathscr{B}\!il(E'_{w*}\times F'_{w*})_+\cap(E\otimes F).$$

Proof. By Remark 10.1, $E_+\otimes^\varepsilon F_+$ is the dual cone of $E'_+\otimes^\pi F'_+$ with respect to the dual pair $\langle E\otimes F, E'\otimes F'\rangle$, so we have $E_+\otimes^\varepsilon F_+ = (E'\otimes F')^*_+\cap(E\otimes F)$. It follows from Proposition 9.1 that $u\in E\otimes F$ belongs to $E_+\otimes^\varepsilon F_+$ if and only if u defines a positive bilinear map $E'\times F'\to\mathbb{R}$. ∎

Corollary 10.5. *All inclusions in* (10.3) *are bipositive.*

Situation II: injective topology, completed

Let E and F be locally convex. Let $E\circledast_\varepsilon F$ denote the space $E\circledast F$ $(=\mathfrak{Bil}(E'_{w*}\times F'_{w*}))$ equipped with the *bi-equicontinuous* (or *injective*) topology ε, that is, the locally convex topology given by the family of seminorms

$$p_{M,N}(b)=\sup_{\varphi\in M,\psi\in N}|b(\varphi,\psi)|,\qquad (M\subseteq E' \text{ and } N\subseteq F' \text{ equicontinuous}).$$

If E and F are complete, then $E\circledast_\varepsilon F$ is also complete (see [Köt79, §40.4.(5)]), so in this case we may identify $E\,\tilde\otimes_\varepsilon F$ with the closure of $E\otimes_\varepsilon F$ in $E\circledast_\varepsilon F$, and we have the following inclusions of vector spaces:

$$E\otimes F\subseteq E\,\tilde\otimes_\varepsilon F\subseteq E\circledast_\varepsilon F\subseteq\mathrm{Bil}(E'\times F'),\qquad (E \text{ and } F \text{ complete}). \tag{10.6}$$

This may fail if E or F is not complete. (In particular, $E\otimes\mathbb{R}=E\circledast\mathbb{R}=E$, but $E\,\tilde\otimes_\varepsilon\mathbb{R}=\tilde E$.) However, in general we have $E\,\tilde\otimes_\varepsilon F=\tilde E\,\tilde\otimes_\varepsilon\tilde F$ (see [Köt79, §44.5.(1)]), hence

$$E\otimes F\subseteq E\,\tilde\otimes_\varepsilon F=\tilde E\,\tilde\otimes_\varepsilon\tilde F\subseteq\tilde E\circledast_\varepsilon\tilde F\subseteq\mathrm{Bil}(E'\times F'). \tag{10.7}$$

Proposition 10.8. *Let E, F be locally convex. Then the natural inclusion $E \tilde{\otimes}_\varepsilon F \hookrightarrow \tilde{E} \circledast_\varepsilon \tilde{F}$ is bipositive; that is:*

$$E_+ \tilde{\otimes}^\varepsilon_\varepsilon \ F_+ = \mathfrak{Bil}\left(E'_{\sigma(E',\tilde{E})} \times F'_{\sigma(F',\tilde{F})}\right)_+ \cap (E \tilde{\otimes}_\varepsilon F).$$

Proof. Continuous linear functionals $\varphi \in E'$ and $\psi \in F'$ define a functional on $E \tilde{\otimes}_\varepsilon F$ in two different ways: either as the (unique) extension of $\varphi \otimes \psi$ to the completion $E \tilde{\otimes}_\varepsilon F$, or as the restriction of the evaluation functional $f_{\varphi,\psi} : \mathrm{Bil}(E' \times F') \to \mathbb{R}$, $b \mapsto b(\varphi, \psi)$ to the subspace $E \tilde{\otimes}_\varepsilon F$. We claim that these two functionals coincide on $E \tilde{\otimes}_\varepsilon F$. The inclusion $E \otimes F \hookrightarrow \mathrm{Bil}(E' \times F')$ is such that $(\varphi \otimes \psi)(u) = u(\varphi, \psi)$, so the functionals coincide on $E \otimes F$. Furthermore, the functional $f_{\varphi,\psi}$ is easily seen to be continuous on $\tilde{E} \circledast_\varepsilon \tilde{F}$ (use that the sets $\{\varphi\} \subseteq E'$, $\{\psi\} \subseteq F'$ are equicontinuous). Hence $\varphi \otimes \psi = f_{\varphi,\psi}$ on $E \otimes F$, and by continuity also on $E \tilde{\otimes}_\varepsilon F$, which proves our claim.

It follows from the claim and the definition of $E_+ \tilde{\otimes}^\varepsilon_\varepsilon \ F_+$ that an element $u \in E \tilde{\otimes}_\varepsilon F$ belongs to $E_+ \tilde{\otimes}^\varepsilon_\varepsilon \ F_+$ if and only if it defines a positive bilinear form $E' \times F' \to \mathbb{R}$. ∎

Corollary 10.9. *All inclusions in* (10.6) *and* (10.7) *are bipositive.*

We only needed the bi-equicontinuous topology on $E \circledast F$ for the proof of Proposition 10.8. From here on out we can forget about it.

Situation III: arbitrary compatible topology, completed

Now let α be an arbitrary compatible topology on $E \otimes F$ (E and F locally convex). Since the injective topology is the weakest compatible topology, we have a natural map $E \tilde{\otimes}_\alpha F \to E \tilde{\otimes}_\varepsilon F$, so here the picture is as follows:

$$E \otimes F \ \hookrightarrow \ E \tilde{\otimes}_\alpha F \ \to \ E \tilde{\otimes}_\varepsilon F \ \hookrightarrow \ \tilde{E} \circledast \tilde{F} \ \hookrightarrow \ \mathrm{Bil}(E' \times F'). \tag{10.10}$$

The map $E \tilde{\otimes}_\alpha F \to E \tilde{\otimes}_\varepsilon F$ need not be injective (this is related to the approximation property; see e.g. [DF93, Theorem 5.6]). However, it remains bipositive.

Proposition 10.11. *Let E, F be locally convex, and let α be a compatible topology on $E \otimes F$. Then the natural map $\Phi_{\alpha\to\varepsilon} : E \tilde{\otimes}_\alpha F \to E \tilde{\otimes}_\varepsilon F$ is bipositive; that is:*

$$E_+ \tilde{\otimes}^\varepsilon_\alpha \ F_+ = \Phi^{-1}_{\alpha\to\varepsilon}[E_+ \tilde{\otimes}^\varepsilon_\varepsilon \ F_+].$$

Proof. Note that $\varphi \tilde{\otimes}_\alpha \psi = (\varphi \tilde{\otimes}_\varepsilon \psi) \circ \Phi_{\alpha\to\varepsilon}$, as they coincide on $E \otimes F$. Hence: $u \in E_+ \tilde{\otimes}^\varepsilon_\alpha \ F_+$ if and only if $\Phi_{\alpha\to\varepsilon}(u) \in E_+ \tilde{\otimes}^\varepsilon_\varepsilon \ F_+$. ∎

Corollary 10.12. *All maps in* (10.10) *are bipositive.*

10.2 Mapping properties of the injective cone

We show that the injective cone preserves all positive maps, bipositive maps (provided the cones are closed), and retracts, and show that it fails to preserve quotients, pushforwards, and approximate pushforwards.

Let $\langle E, E'\rangle$, $\langle F, F'\rangle$, $\langle G, G'\rangle$, $\langle H, H'\rangle$ be dual pairs, equipped with convex cones E_+, F_+, G_+, H_+ in the primal spaces. Given $T \in \mathfrak{L}(E_w, G_w)$ and $S \in \mathfrak{L}(F_w, H_w)$, we define $T \boxtimes S : \mathrm{Bil}(E' \times F') \to \mathrm{Bil}(G' \times H')$ by

$$b \mapsto \Big((\varphi, \psi) \mapsto b(T'\varphi, S'\psi)\Big),$$

where $T' \in \mathfrak{L}(G'_{w*}, E'_{w*})$, $S' \in \mathfrak{L}(H'_{w*}, F'_{w*})$ denote the respective adjoints.

Note that $(T \boxtimes S)b$ is separately weak-$*$ continuous whenever b is, so $T \boxtimes S$ restricts to a map $T \circledast S : E \circledast F \to G \circledast H$.

Proposition 10.13. *The following diagram commutes:*

$$\begin{array}{ccccc} E \otimes F & \hookrightarrow & E \circledast F & \hookrightarrow & \mathrm{Bil}(E' \times F') \\ \downarrow{\scriptstyle T\otimes S} & & \downarrow{\scriptstyle T\circledast S} & & \downarrow{\scriptstyle T\boxtimes S} \\ G \otimes H & \hookrightarrow & G \circledast H & \hookrightarrow & \mathrm{Bil}(G' \times H'). \end{array}$$

Proof. The rightmost square commutes by definition ($T \circledast S$ is the restriction of $T \boxtimes S$). For the leftmost square, note that $x \otimes y \in E \otimes F$ defines the bilinear map $(\varphi, \psi) \mapsto \langle x, \varphi\rangle\langle y, \psi\rangle$, and $Tx \otimes Sy$ defines the bilinear map $(\varphi, \psi) \mapsto \langle Tx, \varphi\rangle\langle Sy, \psi\rangle = \langle x, T'\varphi\rangle\langle y, S'\varphi\rangle$. ∎

Proposition 10.14. *If E, F, G, H are locally convex, if $T \in \mathfrak{L}(E, G)$, $S \in \mathfrak{L}(F, H)$, and if α and β are compatible topologies on $E \otimes F$ and $G \otimes H$ for which the map $T \otimes_{\alpha\to\beta} S : E \otimes_\alpha F \to G \otimes_\beta H$ is continuous, then the following diagram commutes:*

$$\begin{array}{ccccccccc} E \otimes F & \hookrightarrow & E \tilde{\otimes}_\alpha F & \to & E \tilde{\otimes}_\varepsilon F & \hookrightarrow & \tilde{E} \circledast \tilde{F} & \hookrightarrow & \mathrm{Bil}(E' \times F') \\ \downarrow{\scriptstyle T\otimes S} & & \downarrow{\scriptstyle T\tilde{\otimes}_{\alpha\to\beta}S} & & \downarrow{\scriptstyle T\tilde{\otimes}_\varepsilon S} & & \downarrow{\scriptstyle \tilde{T}\circledast\tilde{S}} & & \downarrow{\scriptstyle T\boxtimes S} \\ G \otimes H & \hookrightarrow & G \tilde{\otimes}_\beta H & \to & G \tilde{\otimes}_\varepsilon H & \hookrightarrow & \tilde{G} \circledast \tilde{H} & \hookrightarrow & \mathrm{Bil}(G' \times H'). \end{array}$$

Here the horizontal maps are the ones from (10.10), which are bipositive by Corollary 10.12.

Proof. The rightmost square commutes since $T \boxtimes S = \tilde{T} \boxtimes \tilde{S}$ (use that $T : E \to G$ and its completion $\tilde{T} : \tilde{E} \to \tilde{G}$ have the same adjoint $T' = \tilde{T}' : G' \to E'$), and $\tilde{T} \circledast \tilde{S}$ is a restriction of $\tilde{T} \boxtimes \tilde{S}$. (However, $\tilde{T} \circledast \tilde{S} \neq T \circledast S$, as the domain and codomain are different!)

The other squares (and the triangles) commute because the respective compositions agree on the dense subspace $E \otimes F$ (or $G \otimes H$). ∎

Using the preceding results, we can now show that the injective cone preserves positive maps and approximately bipositive maps.

Lemma 10.15. *Let $\langle E, E'\rangle$, $\langle F, F'\rangle$, $\langle G, G'\rangle$, $\langle H, H'\rangle$ be dual pairs, and let $T \in \mathfrak{L}(E_w, G_w)$ and $S \in \mathfrak{L}(F_w, H_w)$.*

(a) If T and S are positive, then $T \boxtimes S$ is positive.

(b) If $\overline{E_+}^w = T^{-1}[\overline{G_+}^w]$ and $\overline{F_+}^w = S^{-1}[\overline{H_+}^w]$ (i.e. T and S are approximately bipositive), then $T \circledast S$ is bipositive.

Proof.

(a) Let $b \in \mathrm{Bil}(E' \times F')$ be positive. If $\varphi \in G'_+$ and $\psi \in H'_+$, then $\varphi \circ T \in E'_+$ and $\psi \circ S \in F'_+$ (the composition of positive linear maps is positive), so $(T \boxtimes S)(b)(\varphi, \psi) \geq 0$. It follows that $(T \boxtimes S)(b)$ is a positive bilinear map on $G' \times H'$, so $T \boxtimes S$ is positive.

(b) By the duality between approximate pushforwards and approximate pullbacks (see page 111), the adjoints $T' \in \mathfrak{L}(G'_{w*}, E'_{w*})$ and $S' \in \mathfrak{L}(H'_{w*}, F'_{w*})$ are weak-$*$ approximate pushforwards. Since $T \circledast S$ is precisely the map $(T' \otimes S')'$ from Lemma 9.3, it follows from said lemma that $T \circledast S$ is bipositive. ∎

Theorem 10.16. *Let $T \in \mathfrak{L}(E_w, G_w)$ and $S \in \mathfrak{L}(F_w, H_w)$.*

(a) If T and S are positive, then $(T \otimes S)[E_+ \otimes^\varepsilon F_+] \subseteq G_+ \otimes^\varepsilon H_+$.

(b) If $\overline{E_+}^w = T^{-1}[\overline{G_+}^w]$ and $\overline{F_+}^w = S^{-1}[\overline{H_+}^w]$ (i.e. T and S are approximately bipositive), then $E_+ \otimes^\varepsilon F_+ = (T \otimes S)^{-1}[G_+ \otimes^\varepsilon H_+]$.

In summary: the algebraic injective cone preserves continuous positive maps and (continuous[2]) approximately bipositive maps.

Proof. All horizontal arrows in the diagram from Proposition 10.13 are bipositive (by Corollary 10.5), so (a) and (b) follow easily from Lemma 10.15. For the summary, recall from Remark 10.1 that $E_+ \otimes^\varepsilon F_+$ and $G_+ \otimes^\varepsilon H_+$ are weakly closed, so in (b) we find that $T \otimes S$ is approximately bipositive (in addition to being bipositive). ∎

Theorem 10.17. *Let E, F, G, H be locally convex, let $T \in \mathfrak{L}(E, G)$ and $S \in \mathfrak{L}(F, H)$, and let α and β be compatible topologies on respectively $E \otimes F$ and $G \otimes H$ for which the map $T \otimes_{\alpha\to\beta} S : E \otimes_\alpha F \to G \otimes_\beta H$ is continuous.*

(a) If T and S are positive, then $(T \tilde{\otimes}_{\alpha\to\beta} S)[E_+ \tilde{\otimes}^\varepsilon_\alpha F_+] \subseteq G_+ \tilde{\otimes}^\varepsilon_\beta H_+$.

(b) If E and F are complete and $\overline{E_+} = T^{-1}[\overline{G_+}]$ and $\overline{F_+} = S^{-1}[\overline{H_+}]$, then $E_+ \tilde{\otimes}^\varepsilon_\alpha F_+ = (T \tilde{\otimes}_{\alpha\to\beta} S)^{-1}[G_+ \tilde{\otimes}^\varepsilon_\beta H_+]$.

In summary: the completed injective cone preserves continuous positive maps, and (continuous[2]) approximately bipositive maps if E and F are complete.

[2]By our definition, approximately bipositive maps are already required to be continuous.

Proof.

(a) All horizontal arrows in the diagram from Proposition 10.14 are bipositive (by Corollary 10.12), so the result follows from Lemma 10.15(a).

(b) Recall: in a locally convex space, the weak closure and original closure of a convex cone coincide. Moreover, note that we may assume without loss of generality that G and H are also complete. (Extend T to the map $\tilde{T} : E \to \tilde{G}$, and let $\widetilde{G_+}$ denote the closure of G_+ in $\tilde{G}$. Then $\tilde{T}^{-1}[\widetilde{G_+}] = T^{-1}[\overline{G_+}]$, since $\mathrm{ran}(\tilde{T}) \subseteq G$.)

We refer again to the diagram from Proposition 10.14. All horizontal arrows in are bipositive, and the vertical arrow $T \circledast S = \tilde{T} \circledast \tilde{S}$ is bipositive by Lemma 10.15(b). The result is easily deduced. ∎

Remark 10.18. We get one of the characteristic properties of the injective topology for free: if E, F, G, H are locally convex, E and F complete, and if $T \in \mathfrak{L}(E, G)$ and $S \in \mathfrak{L}(F, H)$ are injective, then so is $T \,\tilde{\otimes}_\varepsilon\, S \in \mathfrak{L}(E \,\tilde{\otimes}_\varepsilon\, F, G \,\tilde{\otimes}_\varepsilon\, H)$. Indeed, equip all spaces with the trivial cone $\{0\}$, then every dual cone is the entire dual space, so $\mathrm{Bil}(E' \times F')_+ = \{0\}$. Therefore $E \,\tilde{\otimes}_\varepsilon\, F$ and $G \,\tilde{\otimes}_\varepsilon\, H$ are also equipped with the zero cone (since $E \,\tilde{\otimes}_\varepsilon\, F \to \mathrm{Bil}(E' \times F')$ is bipositive and injective). Since $T \,\tilde{\otimes}_\varepsilon\, S$ is bipositive, we have $(T \,\tilde{\otimes}_\varepsilon\, S)^{-1}[\{0\}] = \{0\}$, so $T \,\tilde{\otimes}_\varepsilon\, S$ is injective.

This shows immediately that the completeness assumptions in Theorem 10.17(b) cannot be omitted. (After all, $T \,\tilde{\otimes}_\varepsilon\, \mathrm{id}_\mathbb{R} : E \,\tilde{\otimes}_\varepsilon\, \mathbb{R} \to G \,\tilde{\otimes}_\varepsilon\, \mathbb{R}$ is simply the completion $\tilde{T} : \tilde{E} \to \tilde{G}$, which may fail to be injective even if T is injective.)

A similar argument shows that the weak closures in Lemma 10.15(b) and subsequent theorems cannot be omitted: the map $T \otimes_\varepsilon \mathrm{id}_\mathbb{R} : E \otimes \mathbb{R} \to G \otimes \mathbb{R}$ is simply T, but with the positive cones E_+, G_+ replaced by their weak closures. But one does not necessarily have $T^{-1}[\overline{G_+}^w] = \overline{E_+}^w$ whenever $T^{-1}[G_+] = E_+$. (Concrete example: let $G = \mathbb{R}^2$ with $G_+ = \{(x, y) : x > 0\} \cup \{(0, 0)\}$, let $E := \mathrm{span}\{(0, 1)\} \subseteq G$ with $E_+ := G_+ \cap E$, and let T be the inclusion $E \hookrightarrow G$.)

Remark 10.19. A topological order retract $G \subseteq E$ is given by two continuous positive linear maps $E \twoheadrightarrow G \hookrightarrow E$, so it follows at once that the injective cone (in all its incarnations) preserves all topological order retracts, without any assumptions on completeness or weak closures. The argument is analogous to that of Proposition 9.2(c).

The following example shows that the injective cone does not preserve pushforwards, not even approximately.

Example 10.20 (Dual to Example 9.7; cf. [Dob22, Situation 4])**.** Let E be a finite-dimensional space equipped with a proper, generating, polyhedral cone which is *not* a simplex cone. Let $x_1, \ldots, x_m$ be representatives of the extremal rays of E_+, and let $\mathbb{R}^m$ be equipped with the standard cone $\mathbb{R}^m_{\geq 0}$. Then the map $T : \mathbb{R}^m \to E$, $(\lambda_1, \ldots, \lambda_m) \mapsto \lambda_1 x_1 + \ldots + \lambda_m x_m$ is a pushforward (i.e. $T[\mathbb{R}^m_{\geq 0}] = E_+$).

Since E_+ is not a simplex cone, it follows from [BL75, Proposition 3.1] (see also Theorem 13.2 below) that $E_+ \otimes^\pi E_+^* \neq E_+ \otimes^\varepsilon E_+^*$. On the other hand, we have

$\mathbb{R}^m_{\geq 0} \otimes^\pi E^*_+ = \mathbb{R}^m_{\geq 0} \otimes^\varepsilon E^*_+$, and it follows from Proposition 9.2(b) that $T \otimes \mathrm{id}_{E^*}$ is a pushforward for the projective cone. Therefore:

$$(T \otimes \mathrm{id}_{E^*})[\mathbb{R}^m_{\geq 0} \otimes^\varepsilon E^*_+] = (T \otimes \mathrm{id}_{E^*})[\mathbb{R}^m_{\geq 0} \otimes^\pi E^*_+] = E_+ \otimes^\pi E^*_+ \neq E_+ \otimes^\varepsilon E^*_+.$$

This shows that $T \otimes \mathrm{id}_{E^*}$ is not a pushforward for the injective cone.

Note that all cones in this example are polyhedral, and therefore closed. In particular, the situation is not resolved by adding closures, which shows that the injective cone does not preserve approximate pushforwards. △

The finite-dimensional techniques used in Example 10.20 will be discussed in more detail in Chapter 12 and Chapter 13.

10.3 When is the injective cone proper?

We determine the lineality space of $E \circledast F$, and we use this to give necessary and sufficient conditions for the injective cone (in all its incarnations) to be proper. Direct formulas for the lineality space (under certain topological assumptions) will be given in §10.5.

As before, let $\langle E, E' \rangle$, $\langle F, F' \rangle$ be dual pairs, equipped with convex cones $E_+ \subseteq E$, $F_+ \subseteq F$ in the primal spaces.

Proposition 10.21. *The lineality space of $(E \circledast F)_+$ is the set of those bilinear forms in $E \circledast F$ that vanish on $\overline{\mathrm{span}(E'_+)}^{w*} \times \overline{\mathrm{span}(F'_+)}^{w*} = \mathrm{lin}(\overline{E_+}^w)^\perp \times \mathrm{lin}(\overline{F_+}^w)^\perp$.*

Proof. If $b \in E \circledast F$ vanishes on $\overline{\mathrm{span}(E'_+)}^{w*} \times \overline{\mathrm{span}(F'_+)}^{w*}$, then in particular it vanishes on $E'_+ \times F'_+$, so evidently both b and $-b$ define positive bilinear forms. Conversely, if $b \in \mathrm{lin}((E \circledast F)_+)$, then both b and $-b$ are positive on $E'_+ \times F'_+$, so it follows that b must vanish on $E'_+ \times F'_+$. Therefore b also vanishes on $\mathrm{span}(E'_+) \times \mathrm{span}(F'_+)$, and consequently on $\overline{\mathrm{span}(E'_+)}^{w*} \times \overline{\mathrm{span}(F'_+)}^{w*}$. (Use weak-$*$ continuity in one variable at a time, as we did in the proof of Lemma 9.3.)

Since $\mathrm{lin}(\overline{E_+}^w) = {}^\perp(E'_+)$ (see §8.3), we have $\overline{\mathrm{span}(E'_+)}^{w*} = \mathrm{lin}(\overline{E_+}^w)^\perp$. ■

Direct formulas for the lineality space of the injective cone will be given in Corollary 10.37(c) (in $E \otimes F$) and Corollary 10.41(b) (in $E \circledast F$). For now, we focus on conditions for the injective cone to be proper.

Theorem 10.22. *The following are equivalent:*

(i) $E_+ \otimes^\varepsilon F_+$ is a proper cone;

(ii) For every subspace $E \otimes F \subseteq G \subseteq E \circledast F$, the cone $G_+ := G \cap (E \circledast F)_+$ is proper.

(iii) $(E \circledast F)_+$ is a proper cone;

(iv) $E = \{0\}$, or $F = \{0\}$, or both $\overline{E_+}^w$ and $\overline{F_+}^w$ are proper cones.

In particular, the injective tensor product of weakly closed proper cones is a proper cone.

Note that the equivalence *(i)* $\iff$ *(iv)* is very similar to Theorem 9.10. However, we should point out that the corner case is slightly different now. In Theorem 9.10, the corner case is when one of the *cones* is trivial; here the corner case is when one of the *spaces* is trivial.

Proof of Theorem 10.22. *(iii)* $\Longrightarrow$ *(ii)*. Trivial.

(ii) $\Longrightarrow$ *(i)*. Immediate, since $E_+ \otimes^\varepsilon F_+ = (E \otimes F) \cap (E \circledast F)_+$.

(iv) $\Longrightarrow$ *(iii)*. If $E = \{0\}$, then clearly $E \circledast F = \{0\}$, so $(E \circledast F)_+$ is a proper cone regardless of any properties of F_+ (and similarly if $F = \{0\}$). If $\overline{E_+}^w$ and $\overline{F_+}^w$ are proper cones, then $\lin(\overline{E_+}^w) = \lin(\overline{F_+}^w) = \{0\}$, so it follows from Proposition 10.21 that $\lin((E \circledast F)_+) = \{0\}$.

(i) $\Longrightarrow$ *(iv)*. We prove the contrapositive: suppose that $E, F \neq \{0\}$ and that $\overline{E_+}^w$ is not a proper cone. Then we may choose $x \in E \setminus \{0\}$ with $\pm x \in \overline{E_+}^w$. Note that $(\overline{E_+}^w)' = E_+'$, so for every $\varphi \in E_+'$ we have $\varphi(x), \varphi(-x) \geq 0$, and therefore $\varphi(x) = 0$. Now choose any $y \in F \setminus \{0\}$ (here we use that $F \neq \{0\}$), then for all $\varphi \in E_+'$, $\psi \in F_+'$ we have $\langle x \otimes y, \varphi \otimes \psi \rangle = \varphi(x)\psi(y) = 0 \cdot \psi(y) = 0$, so we find $\pm x \otimes y \in E_+ \otimes^\varepsilon F_+$. Since x and y are non-zero, we have $x \otimes y \neq 0$, and we conclude that $E_+ \otimes^\varepsilon F_+$ fails to be proper. ∎

To tell whether $E_+ \tilde{\otimes}_\alpha^\varepsilon F_+$ is a proper cone, we need to assume that E and F are complete. In the case where E and F are not complete, an answer can be found by first passing to the completions $\tilde{E}$, $\tilde{F}$. See also Remark 10.24 below.

Corollary 10.23. *Let E, F be complete locally convex spaces, $E_+ \subseteq E$, $F_+ \subseteq F$ convex cones, and α a compatible locally convex topology on $E \otimes F$. Then the following are equivalent:*

(i) $E_+ \tilde{\otimes}_\alpha^\varepsilon F_+ \subseteq E \tilde{\otimes}_\alpha F$ *is a proper cone;*

(ii) $E = \{0\}$*, or* $F = \{0\}$*, or both* $\overline{E_+}$ *and* $\overline{F_+}$ *are proper cones and the natural map* $E \tilde{\otimes}_\alpha F \to E \tilde{\otimes}_\varepsilon F$ *is injective.*

Proof. First of all, recall that $\overline{E_+} = \overline{E_+}^w$, since E_+ is convex and E is locally convex. Likewise, $\overline{F_+} = \overline{F_+}^w$.

For the injective topology, recall from (10.6) that we have $E \otimes F \subseteq E \tilde{\otimes}_\varepsilon F \subseteq E \circledast F$, since E and F are complete. Hence for $\alpha = \varepsilon$ the result follows from Theorem 10.22.

For general α, recall that $E \tilde{\otimes}_\alpha F \to E \tilde{\otimes}_\varepsilon F$ is bipositive. Therefore:

(i) $\Longrightarrow$ *(ii)*. If $E_+ \tilde{\otimes}_\alpha^\varepsilon F_+$ is proper, then the bipositive map $E \tilde{\otimes}_\alpha F \to E \tilde{\otimes}_\varepsilon F$ is automatically injective (see Remark A.7). Furthermore, the subcone $E_+ \otimes^\varepsilon F_+ \subseteq E_+ \tilde{\otimes}_\alpha^\varepsilon F_+$ is also proper, so it follows from Theorem 10.22 that *(ii)* holds.

(ii) $\Longrightarrow$ *(i)*. It follows from the assumptions that $E_+ \tilde{\otimes}_\varepsilon^\varepsilon F_+$ is a proper cone and that $E \tilde{\otimes}_\alpha F \to E \tilde{\otimes}_\varepsilon F$ is injective. (The latter statement is trivially true if $E = \{0\}$ or $F = \{0\}$; otherwise it holds by assumption.) Since $E \tilde{\otimes}_\alpha F \to E \tilde{\otimes}_\varepsilon F$ is bipositive and injective, it follows that $E_+ \tilde{\otimes}_\alpha^\varepsilon F_+$ is also proper. ∎

Remark 10.24. In Corollary 10.23, the assumption that E and F are complete cannot be omitted. Under the natural isomorphism $E \tilde{\otimes}_\alpha \mathbb{R} \cong \tilde{E}$, the injective cone

$E_+ \mathbin{\tilde{\otimes}^{\varepsilon}_{\alpha}} \mathbb{R}_+$ corresponds with $\widetilde{E_+}$ (the closure of E_+ in $\tilde{E}$). However, it can happen that $\overline{E_+}$ is proper but $\widetilde{E_+}$ is not (e.g. [Dob20a, Example 6.4]).

Remark 10.25. The natural map $E \mathbin{\tilde{\otimes}_{\alpha}} F \to E \mathbin{\tilde{\otimes}_{\varepsilon}} F$ is not always injective; this is related to the approximation property. Further remarks along this line can be found in §11.4 below; see also [DF93, Theorem 5.6].

10.4 Faces of the injective cone

In this section, we present a general way to construct faces of the space $E \circledast F = \mathfrak{Bil}(E'_{w*} \times F'_{w*})$ of separately weak-$*$ continuous bilinear forms. This will be used in §10.5 to obtain ideals in for the injective cone.

Since the injective cone is characterized by bipositive maps $E \otimes F \to E \circledast F$ and $E \mathbin{\tilde{\otimes}_{\alpha}} F \to \tilde{E} \circledast \tilde{F}$ (see §10.1, the inverse image of a face in $(E \circledast F)_+$ (resp. $(\tilde{E} \circledast \tilde{F})_+$) immediately gives us a face in $E_+ \otimes^{\varepsilon} F_+$ (resp. $E_+ \mathbin{\tilde{\otimes}^{\varepsilon}_{\alpha}} F_+$). Therefore we focus on faces in $E \circledast F$. For ideals in $E \otimes F$ and $E \mathbin{\tilde{\otimes}_{\alpha}} F$, see §10.5.

Definition 10.26. Let $\langle E, E' \rangle$, $\langle F, F' \rangle$ be dual pairs, and let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones. Given $b \in E \circledast F$ and subsets $M' \subseteq E'$, $N' \subseteq F'$, let us write

$$b(M', \cdot) := \{b(\varphi, \cdot) \,:\, \varphi \in M'\} \subseteq (F'_{w*})' = F;$$
$$b(\cdot, N') := \{b(\cdot, \psi) \,:\, \psi \in N'\} \subseteq (E'_{w*})' = E.$$

Given subsets $M \subseteq E$, $M' \subseteq E'$, $N \subseteq F$, $N' \subseteq F'$, we define

$$M' \ltimes N := \big\{b \in E \circledast F \,:\, b(M', \cdot) \subseteq N\big\};$$
$$M \rtimes N' := \big\{b \in E \circledast F \,:\, b(\cdot, N') \subseteq M\big\}.$$

Under the natural isomorphism $E \circledast F = \mathfrak{Bil}(E'_{w*} \times F'_{w*}) \cong \mathfrak{L}(E'_{w*}, F_w)$, the set $M' \ltimes N$ is simply the set of operators $T : E'_{w*} \to F_w$ satisfying $T[M'] \subseteq N$. Likewise, $M \rtimes N'$ corresponds with the set of operators $S : F'_{w*} \to E_w$ satisfying $S[N'] \subseteq M$.

Note that the positive cone can be described as $(E \circledast F)_+ = E'_+ \ltimes \overline{F_+}^{\,w} = \overline{E_+}^{\,w} \rtimes F'_+$.

The following lemma will be central to the remainder of this chapter. The special case where M' and N' are also faces has already been studied in the finite-dimensional setting (e.g. [Bar78b, §4] and [Tam92, §4]), but we will need the greater generality presented here.

Lemma 10.27. *If $M' \subseteq E'_+$, $N' \subseteq F'_+$ are subsets of the dual cones and if $M \subseteq \overline{E_+}^{\,w}$, $N \subseteq \overline{F_+}^{\,w}$ are faces, then $(M' \ltimes N) \cap (E \circledast F)_+$ and $(M \rtimes N') \cap (E \circledast F)_+$ are faces of $(E \circledast F)_+$.*

Proof. Given $\varphi \in E'$, let $L_\varphi : E \circledast F \to (F'_{w*})' = F$ denote the map $b \mapsto b(\varphi, \cdot)$. If $\varphi \in E'_+$, then L_φ is a positive linear map in the sense that $L_\varphi[(E \circledast F)_+] \subseteq \overline{F_+}^{\,w}$.

Therefore $L_\varphi^{-1}[N] \cap (E \circledast F)_+$ defines a face of $(E \circledast F)_+$. Since $(M' \ltimes N) \cap (E \circledast F)_+$ can be written as an intersection of faces,

$$(M' \ltimes N) \cap (E \circledast F)_+ \;=\; \bigcap_{\varphi \in M'} L_\varphi^{-1}[N] \cap (E \circledast F)_+,$$

it also a face of $(E \circledast F)_+$. The conclusion for $(M \rtimes N') \cap (E \circledast F)_+$ follows by symmetry. ■

As a first application of Lemma 10.27, we study a construction of faces in the injective cone that is dual to the construction in the projective cone (see §9.4). A slightly different construction, based again on Lemma 10.27, will be used in §10.5 below to construct ideals for the injective cone.

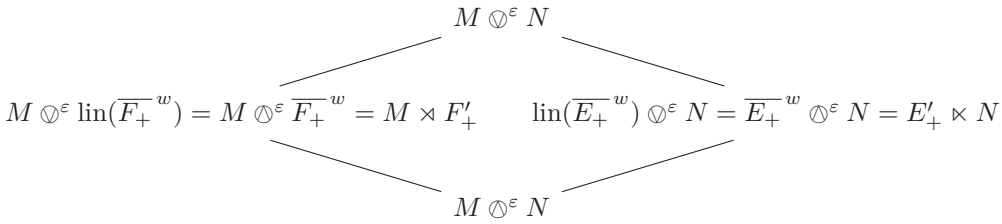
Theorem 10.28. *Let $M \subseteq \overline{E_+}^w$, $N \subseteq \overline{F_+}^w$ be faces, and define*

$$M \mathbin{\textcircled{\scriptsize\vee}}^\varepsilon N := (M \rtimes N^\diamond) \cap (M^\diamond \ltimes N) \cap (E \circledast F)_+;$$
$$M \mathbin{\textcircled{\scriptsize\wedge}}^\varepsilon N := (M \rtimes F'_+) \cap (E'_+ \ltimes N).$$

Then:

(a) $M \mathbin{\textcircled{\scriptsize\vee}}^\varepsilon N$ and $M \mathbin{\textcircled{\scriptsize\wedge}}^\varepsilon N$ are faces of $(E \circledast F)_+$.

(b) The face lattice of $(E \circledast F)_+$ contains the following partially ordered subset:



This subset respects meets from the face lattice:

$$M \mathbin{\textcircled{\scriptsize\wedge}}^\varepsilon N = (M \mathbin{\textcircled{\scriptsize\vee}}^\varepsilon \operatorname{lin}(\overline{F_+}^w)) \cap (\operatorname{lin}(\overline{E_+}^w) \mathbin{\textcircled{\scriptsize\vee}}^\varepsilon N) = (M \mathbin{\textcircled{\scriptsize\wedge}}^\varepsilon \overline{F_+}^w) \cap (\overline{E_+}^w \mathbin{\textcircled{\scriptsize\wedge}}^\varepsilon N).$$

(c) If M and N are dual faces, then so are $M \mathbin{\textcircled{\scriptsize\vee}}^\varepsilon N$ and $M \mathbin{\textcircled{\scriptsize\wedge}}^\varepsilon N$, and one has

$$M \mathbin{\textcircled{\scriptsize\vee}}^\varepsilon N = {}^\diamond(M^\diamond \mathbin{\textcircled{\scriptsize\wedge}}^\pi N^\diamond) = (M \rtimes N^\diamond) \cap (E \circledast F)_+ = (M^\diamond \ltimes N) \cap (E \circledast F)_+;$$
$$M \mathbin{\textcircled{\scriptsize\wedge}}^\varepsilon N = {}^\diamond(M^\diamond \mathbin{\textcircled{\scriptsize\vee}}^\pi N^\diamond).$$

If this is the case, then the subset from (b) respects meets and joins from the lattice of $\langle (E \circledast F)_+ , E'_+ \otimes^\pi F'_+ \rangle$-dual faces (as defined in Appendix A.3).

(d) If M and N are exposed faces, then so is $M \mathbin{\textcircled{\scriptsize\vee}}^\varepsilon N$.

(e) If M and N as well as $\operatorname{lin}(\overline{E_+}^w)$ and $\operatorname{lin}(\overline{F_+}^w)$ are exposed faces, then so is $M \mathbin{\textcircled{\scriptsize\wedge}}^\varepsilon N$.

Note: in the finite-dimensional case, the conclusion in (c) is simply that the four-element subset from (b) respects the operations of the lattice of exposed faces. (Here we use that $(E_+ \otimes^\varepsilon F_+)^* = E_+^* \otimes^\pi F_+^*$ because $E_+^* \otimes^\pi F_+^*$ is closed; see Corollary 12.13(b).)

Proof of Theorem 10.28.

(a) Note that everything in $M \rtimes F'_+$ is automatically positive, for if $b(\,\cdot\,, F'_+) \subseteq M$ then certainly $b(\,\cdot\,, F'_+) \subseteq \overline{E_+}^w$. This shows that $M \rtimes F'_+ = (M \rtimes F'_+) \cap (E \circledast F)_+$. Now the result follows from Lemma 10.27, since the intersection of two faces is again a face.

(b) If $b \in M \rtimes F'_+$, then $b(\,\cdot\,, F'_+) \subseteq M$, so in particular b vanishes on $M^\diamond \times F'_+$. Therefore $b(M^\diamond, \,\cdot\,) \subseteq {}^\perp(F'_+) = \operatorname{lin}(\overline{F_+}^w)$, which shows that $M \rtimes F'_+ \subseteq M^\diamond \ltimes \operatorname{lin}(\overline{F_+}^w)$. Since we also have $M \rtimes F'_+ \subseteq (E \circledast F)_+$ (see (a)), it follows from the definition that

$$M \mathbin{\textcircled{\scriptsize$\vee$}}^\varepsilon \operatorname{lin}(\overline{F_+}^w) = (M \rtimes F'_+) \cap (M^\diamond \ltimes \operatorname{lin}(\overline{F_+}^w)) \cap (E \circledast F)_+ = M \rtimes F'_+.$$

Similarly, since $E'_+ \ltimes \overline{F_+}^w = (E \circledast F)_+$, it follows again from the definition that

$$M \mathbin{\textcircled{\scriptsize$\wedge$}}^\varepsilon \overline{F_+}^w = (M \rtimes F'_+) \cap (E'_+ \ltimes \overline{F_+}^w) \cap (E \circledast F)_+ = M \rtimes F'_+.$$

The equality $\operatorname{lin}(\overline{E_+}^w) \mathbin{\textcircled{\scriptsize$\vee$}}^\varepsilon N = \overline{E_+}^w \mathbin{\textcircled{\scriptsize$\wedge$}}^\varepsilon N = E'_+ \ltimes N$ follows analogously. As a consequence, the intersection formula follows immediately from the definition of $M \mathbin{\textcircled{\scriptsize$\wedge$}}^\varepsilon N$. Finally, the upwards inclusions follow by noting that if $M_1 \subseteq M_2 \subseteq \overline{E_+}^w$ and $N_1 \subseteq N_2 \subseteq \overline{F_+}^w$ are faces, then $M_1 \mathbin{\textcircled{\scriptsize$\vee$}}^\varepsilon N_1 \subseteq M_2 \mathbin{\textcircled{\scriptsize$\vee$}}^\varepsilon N_2$.

(c) If $b \in M \rtimes N^\diamond$, then $b(\,\cdot\,, N^\diamond) \subseteq M$, so in particular b vanishes on $M^\diamond \times N^\diamond$. Conversely, if $b \in (E \circledast F)_+$ vanishes on $M^\diamond \times N^\diamond$, then $b(\,\cdot\,, N^\diamond) \subseteq {}^\diamond(M^\diamond) = M$, so $b \in M \rtimes N^\diamond$. This proves that

$$\begin{aligned}(M \rtimes N^\diamond) \cap (E \circledast F)_+ &= \big\{b \in (E \circledast F)_+ \,:\, b(\varphi, \psi) = 0 \text{ for all } \varphi \in M^\diamond,\, \psi \in N^\diamond\big\} \\ &= {}^\diamond(M^\diamond \otimes_s N^\diamond).\end{aligned}$$

By symmetry, the same is true of $M^\diamond \ltimes N$, so we find

$$M \mathbin{\textcircled{\scriptsize$\vee$}}^\varepsilon N = (M \rtimes N^\diamond) \cap (E \circledast F)_+ = (M^\diamond \ltimes N) \cap (E \circledast F)_+ = {}^\diamond(M^\diamond \otimes_s N^\diamond).$$

Since $M^\diamond \mathbin{\textcircled{\scriptsize$\wedge$}}^\pi N^\diamond$ is the face (of $E'_+ \otimes^\pi F'_+$) generated by $M^\diamond \otimes_s N^\diamond$, it follows that $M \mathbin{\textcircled{\scriptsize$\vee$}}^\varepsilon N = {}^\diamond(M^\diamond \mathbin{\textcircled{\scriptsize$\wedge$}}^\pi N^\diamond)$. This shows that $M \mathbin{\textcircled{\scriptsize$\vee$}}^\varepsilon N$ is a dual face.

Since $\operatorname{lin}(\overline{E_+}^w) = {}^\diamond(E'_+)$ and $\operatorname{lin}(\overline{F_+}^w) = {}^\diamond(F'_+)$ are dual faces, it follows from the intersection formula from (b) that $M \mathbin{\textcircled{\scriptsize$\wedge$}}^\varepsilon N$ is also a dual face. Furthermore, since $\operatorname{lin}(\overline{E_+}^w)^\diamond = E'_+$ and $\operatorname{lin}(\overline{F_+}^w)^\diamond = F'_+$, it follows that

$$\begin{aligned}M \mathbin{\textcircled{\scriptsize$\wedge$}}^\varepsilon N &= (M \mathbin{\textcircled{\scriptsize$\vee$}}^\varepsilon \operatorname{lin}(\overline{F_+}^w)) \cap (\operatorname{lin}(\overline{E_+}^w) \mathbin{\textcircled{\scriptsize$\vee$}}^\varepsilon N) \\ &= {}^\diamond(M^\diamond \mathbin{\textcircled{\scriptsize$\wedge$}}^\pi F'_+) \cap {}^\diamond(E'_+ \mathbin{\textcircled{\scriptsize$\wedge$}}^\pi N^\diamond) \\ &= {}^\diamond\big((M^\diamond \mathbin{\textcircled{\scriptsize$\wedge$}}^\pi F'_+) + (E'_+ \mathbin{\textcircled{\scriptsize$\wedge$}}^\pi N^\diamond)\big) \\ &= {}^\diamond(M^\diamond \mathbin{\textcircled{\scriptsize$\vee$}}^\pi N^\diamond),\end{aligned}$$

where the last step uses that $M^\diamond \mathbin{\circledcirc}^\pi N^\diamond = (M^\diamond \mathbin{\circledcirc}^\pi F'_+) + (E'_+ \mathbin{\circledcirc}^\pi N^\diamond)$, by Theorem 9.13(b).

That the diagram from (b) respects joins from the lattice of $\langle (E \circledast F)_+, E'_+ \otimes^\pi F'_+ \rangle$-dual faces follows from duality. Indeed, by Theorem 9.13(c) and Theorem 9.13(d), $M^\diamond \mathbin{\circledcirc}^\pi N^\diamond$ and $M^\diamond \mathbin{\circledcirc}^\pi N^\diamond$ are $\langle E'_+ \otimes^\pi F'_+, (E \circledast F)_+ \rangle$-dual faces (use that $\mathrm{lin}(E'_+)$ and $\mathrm{lin}(F'_+)$ are automatically dual faces, because E'_+ and F'_+ are weak-$*$ closed; see Remark 9.15), so it follows that

$$(M \mathbin{\circledcirc}^\varepsilon N)^\diamond = M^\diamond \mathbin{\circledcirc}^\pi N^\diamond;$$
$$(M \mathbin{\circledcirc}^\varepsilon N)^\diamond = M^\diamond \mathbin{\circledcirc}^\pi N^\diamond.$$

Therefore, the join of $M \mathbin{\circledcirc}^\varepsilon \mathrm{lin}(\overline{F_+}^w)$ and $\mathrm{lin}(\overline{E_+}^w) \mathbin{\circledcirc}^\varepsilon N$ in the lattice of $\langle (E \circledast F)_+, E'_+ \otimes^\pi F'_+ \rangle$-dual faces is given by

$$\begin{aligned} {}^\diamond\Big(\big(M \mathbin{\circledcirc}^\varepsilon \mathrm{lin}(\overline{F_+}^w)\big)^\diamond \cap \big(\mathrm{lin}(\overline{E_+}^w) \mathbin{\circledcirc}^\varepsilon N\big)^\diamond \Big) &= {}^\diamond\big((M^\diamond \mathbin{\circledcirc}^\pi F'_+) \cap (E'_+ \mathbin{\circledcirc}^\pi N^\diamond)\big) \\ &= {}^\diamond\big(M^\diamond \mathbin{\circledcirc}^\pi N^\diamond\big) \\ &= M \mathbin{\circledcirc}^\varepsilon N. \end{aligned}$$

(d) Suppose that $M = {}^\diamond\{\varphi_0\}$ and $N = {}^\diamond\{\psi_0\}$. Then in particular M and N are dual faces, so by (c) we have

$$M \mathbin{\circledcirc}^\varepsilon N = {}^\diamond(M^\diamond \mathbin{\circledcirc}^\pi N^\diamond) = (M \rtimes N^\diamond) \cap (E \circledast F)_+ = (M^\diamond \ltimes N) \cap (E \circledast F)_+.$$

We prove that $M \mathbin{\circledcirc}^\varepsilon N = {}^\diamond\{\varphi_0 \otimes \psi_0\}$. Evidently one has $\{\varphi_0 \otimes \psi_0\} \subseteq M^\diamond \mathbin{\circledcirc}^\pi N^\diamond$, so ${}^\diamond\{\varphi_0 \otimes \psi_0\} \supseteq {}^\diamond(M^\diamond \mathbin{\circledcirc}^\pi N^\diamond) = M \mathbin{\circledcirc}^\varepsilon N$. For the converse, suppose that $b \in (E \circledast F)_+$ is such that $b(\varphi_0, \psi_0) = 0$. Then $b(\,\cdot\,, \psi_0) \in {}^\diamond\{\varphi_0\} = M$, so b vanishes on $M^\diamond \times \{\psi_0\}$. It follows that $b(M^\diamond, \cdot\,) \subseteq {}^\diamond\{\psi_0\} = N$, so $b \in (M^\diamond \ltimes N) \cap (E \circledast F)_+ = M \mathbin{\circledcirc}^\varepsilon N$.

(e) This follows from (d) and the intersection formula from (b). ■

Remark 10.29. In Theorem 10.28(e), it is required that $\mathrm{lin}(\overline{E_+}^w)$ and $\mathrm{lin}(\overline{F_+}^w)$ are exposed. Recall that this is automatically the case if E and F are separable normed spaces; see Remark 9.15 and Corollary A.20.

Much as in the projective case, this assumption on $\mathrm{lin}(\overline{E_+}^w)$ and $\mathrm{lin}(\overline{F_+}^w)$ cannot be omitted. The example runs along the same lines as the example in Remark 9.15, except we need a much larger space. Let E_+ be a weakly closed proper cone for which $\{0\}$ is not exposed (see Example A.21, Example A.22), and let $F := \mathbb{R}$ with the standard cone, so that $E \circledast F \cong E$. Take some exposed face $M \subseteq E_+$, and let $N := \{0\} \subseteq \mathbb{R}$ be the minimal face. Then $M \mathbin{\circledcirc}^\varepsilon N = \{0\}$, which is not exposed by assumption.

Remark 10.30. Theorem 10.28(c) presents a duality between the four-element sublattices from Theorem 9.13(b) and Theorem 10.28(b). In the projective diagram, the top face $M \mathbin{\circledcirc}^\pi N$ is not merely the join, but even the sum of the left and right faces

$M \between^{\pi} F_+$ and $E_+ \between^{\pi} N$. Given that the injective diagram is dual to the projective diagram, could the same be true here?

Unfortunately, this is not the case, and it already fails for proper, generating, polyhedral cones in finite-dimensional spaces. In this setting, all faces are exposed, so by Theorem 10.28(c) an equivalent question is the following: if $f : E^* \to F$ is positive with $f[M^\diamond] \subseteq N$, then can f be written as $f = g + h$ with g and h positive and $g[M^\diamond] = \{0\}$ and $h[E_+^*] \subseteq N$?

Counterexample: let F_+ be a proper, generating, polyhedral cone with a facet $N \subseteq F_+$ such that at least two extremal rays of F_+ are not contained in N. Furthermore, let $E_+ := F_+^*$ with $M := N^\diamond$, and let $f : E^* = F \to F$ be the identity. Then one has $f[M^\diamond] \subseteq N$. However, if $f = g + h$ is the desired decomposition, then $\operatorname{rank}(g) \leq 1$, because $\ker(g)$ contains a facet, so $g[F_+]$ is either a ray or $\{0\}$. But now every $x \in F_+$ can be written as $x = g(x) + h(x) \in g[F_+] + N$, contradicting our assumption that at least two extremal rays of F_+ are not contained in N.

10.5 Order ideals for the injective cone

Recall that $I \mapsto I_+$ defines a surjective many-to-one correspondence between order ideals and faces (see Appendix A.1). In order to get more convenient formulas for the faces of the injective cone, it is helpful to formulate these results in terms of ideals. The main aim in this section is to provide sufficient conditions so that $I \otimes J$ and $(I \otimes F) + (E \otimes J)$ are ideals for the injective cone, given that $I \subseteq E$ and $J \subseteq F$ are ideals in the base spaces. (Similar questions in $E \circledast F$ and $E \mathbin{\tilde{\otimes}_\alpha} F$ are also addressed.)

Recall from §10.1 that the injective cone is characterized by bipositive maps $E \otimes F \hookrightarrow E \circledast F$ and $E \mathbin{\tilde{\otimes}_\alpha} F \to \tilde{E} \circledast \tilde{F}$. Given subsets $X \subseteq E \circledast F$ and $Y \subseteq \tilde{E} \circledast \tilde{F}$, we denote by $X \cap (E \otimes F)$ and $Y \cap (E \mathbin{\tilde{\otimes}_\alpha} F)$ the inverse images of X and Y under these maps. (This is a slight abuse of notation, for the map $E \mathbin{\tilde{\otimes}_\alpha} F \to \tilde{E} \circledast \tilde{F}$ might fail to be injective in the absence of the approximation property, but this will cause no confusion.) It is not hard to see that the inverse image of an ideal (resp. face) under a bipositive map is again an ideal (reps. face) (see Proposition A.3(b)), so $(E \otimes F) \cap X$ and $(E \mathbin{\tilde{\otimes}_\alpha} F) \cap Y$ are ideals (resp. faces) whenever X and Y are ideals (resp. faces). This is the approach that we will take: we establish ideals in $E \circledast F$ and restrict these to ideals in the algebraic/completed tensor product.

In order to obtain ideals in $E \circledast F$, we note that the faces obtained in Lemma 10.27 can sometimes be written as the positive part of a linear subspace.

Lemma 10.31. *In the notation from §10.4:*

(a) If $M' \subseteq E'$ and $N' \subseteq F'$ are subsets and if $M \subseteq E$ and $N \subseteq F$ are linear subspaces, then $M' \ltimes N$ and $M \rtimes N'$ are linear subspaces.

(b) If $I \subseteq E$ and $J \subseteq F$ are subspaces and if I is weakly closed, then $I^\perp \ltimes J \subseteq I \rtimes J^\perp$.

(c) If $I \subseteq E$ and $J \subseteq F$ are weakly closed subspaces, then $I^\perp \ltimes J = I \rtimes J^\perp = {}^\perp(I^\perp \otimes J^\perp)$, where the orthogonal complement is taken with respect to the dual pair $\langle E \circledast F, E' \otimes F' \rangle$.

Note that ${}^{\perp}(I^{\perp} \otimes J^{\perp}) \subseteq E \circledast F$ is the set of separately weak-$*$ continuous bilinear forms $E' \times F' \to \mathbb{R}$ that vanish on $I^{\perp} \times J^{\perp}$.

Proof of Lemma 10.31.

(a) If $T_1, T_2 : E'_{w*} \to F_w$ map the subset $M' \subseteq E'$ in the subspace $N \subseteq F$, and if $\lambda, \mu \in \mathbb{R}$ are arbitrary, then $\lambda T_1 + \mu T_2$ also maps M' in N.

(b) If $b(I^{\perp}, \cdot\,) \subseteq J$, then $b(I^{\perp}, J^{\perp}) = \{0\}$, hence $b(\,\cdot\,, J^{\perp}) \subseteq {}^{\perp}(I^{\perp}) = I$, since I is weakly closed.

(c) Since J is weakly closed, one has $b(I^{\perp}, \cdot\,) \subseteq J$ if and only if $b(I^{\perp}, J^{\perp}) = \{0\}$, i.e. b vanishes on $I^{\perp} \times J^{\perp}$. Therefore $I^{\perp} \ltimes J = {}^{\perp}(I^{\perp} \otimes J^{\perp})$. The other equality follows analogously. ■

We can now formulate the following "linearization" of Lemma 10.27.

Lemma 10.32. *Let $M' \subseteq E'_+$ be a set of positive linear functionals, and let $N \subseteq \overline{F_+}^{\,w}$ be a face.*

(a) If $J \subseteq F$ is a weakly closed subspace such that $J \cap \overline{F_+}^{\,w} = N$, then

$$(M' \ltimes N) \cap (E \circledast F)_+ \;=\; \big(\overline{\operatorname{span}(M')}^{\,w*} \ltimes J\big) \cap (E \circledast F)_+.$$

In particular, $\overline{\operatorname{span}(M')}^{\,w*} \ltimes J$ is an ideal in $E \circledast F$.

(b) If $J \subseteq F$ is a subspace such that $J \cap \overline{F_+}^{\,w} = N$, then

$$(M' \ltimes N) \cap (E \circledast F)_+ \cap (E \otimes F) \;=\; \big(\overline{\operatorname{span}(M')}^{\,w*} \ltimes J\big) \cap (E \circledast F)_+ \cap (E \otimes F).$$

In particular, $\big(\overline{\operatorname{span}(M')}^{\,w*} \ltimes J\big) \cap (E \otimes F)$ is an ideal in $E \otimes F$.

Interchanging E and F yields corresponding statements for ideals of the form $I \rtimes \overline{\operatorname{span}(N')}^{\,w*}$ and $(I \rtimes \overline{\operatorname{span}(N')}^{\,w*}) \cap (E \otimes F)$.

Proof.

(a) "$\subseteq$". If $b \in M' \ltimes N$, then we have $b(M', \cdot\,) \subseteq N \subseteq J$, so it follows by linearity and continuity that $b(\overline{\operatorname{span}(M')}^{\,w*}, \cdot\,) \subseteq J$. This shows that $M' \ltimes N \subseteq \overline{\operatorname{span}(M')}^{\,w*} \ltimes J$.

"$\supseteq$". If $b \in \big(\overline{\operatorname{span}(M')}^{\,w*} \ltimes J\big) \cap (E \circledast F)_+$, then $b(M', \cdot\,) \subseteq b(\overline{\operatorname{span}(M')}^{\,w*}, \cdot\,) \subseteq J$, but also $b(M', \cdot\,) \subseteq b(E'_+, \cdot\,) \subseteq \overline{F_+}^{\,w}$ by positivity, so we find $b(M', \cdot\,) \subseteq J \cap \overline{F_+}^{\,w} = N$.

To conclude that $\overline{\operatorname{span}(M')}^{\,w*} \ltimes J$ is an ideal, note that it is a linear subspace (by Lemma 10.31(a)) whose positive part is a face (by Lemma 10.27).

(b) "$\subseteq$". If $b \in (M' \ltimes N) \cap (E \otimes F)$, then $b(M', \cdot) \subseteq N \subseteq J$. But b has finite rank, so there is a finite-dimensional (hence closed) subspace $Y \subseteq J$ such that $b(M', \cdot) \subseteq Y$. By linearity and continuity, it follows that $b(\overline{\operatorname{span}(M')}^{w*}, \cdot) \subseteq Y \subseteq J$, which shows that $(M' \ltimes N) \cap (E \otimes F) \subseteq \overline{\operatorname{span}(M')}^{w*} \ltimes J$.

The reverse inclusion "$\supseteq$" and the conclusion follow as in (a). ■

Recall that we call a convex cone $E_+ \subseteq E$ in a topological vector space *semisimple* if $\overline{E_+}^w$ is a proper cone, or equivalently, if $\operatorname{span}(E'_+)$ is weak-$*$ dense in E' (see §8.3 and [Dob20a]). Furthermore, if $I \subseteq E$ is a weakly closed subspace, then the quotient E/I belongs to the dual pair $\langle E/I, I^\perp \rangle$, the weak topology of E/I coincides with the quotient topology E_w/I, and the weak-$*$ topology on $(E/I)' = I^\perp \subseteq E'$ coincides with the relative $\sigma(E', E)$-topology (see §8.2). In particular, we may unambiguously refer to this as the weak-$*$ topology on $(E/I)' \cong I^\perp$.

Theorem 10.33. *Let $\langle E, E' \rangle$, $\langle F, F' \rangle$ be dual pairs, and let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones. Given subspaces $I \subseteq E$ and $J \subseteq F$, we define*

$$I \circledcirc J := (I^\perp \ltimes J) \cap (I \rtimes J^\perp);$$
$$I \circledcirc J := (\operatorname{lin}(\overline{E_+}^w)^\perp \ltimes J) \cap (I \rtimes \operatorname{lin}(\overline{F_+}^w)^\perp).$$

Suppose that I and J are ideals with respect to $\overline{E_+}^w$ and $\overline{F_+}^w$, respectively.[3] *Then:*

(a) $(I \circledcirc J) \cap (E \otimes F)$ is an ideal in $E \otimes F$ (with respect to the injective cone);

(b) If I and J are weakly closed, then $I \circledcirc J$ is an ideal in $E \circledast F$;

(c) If I is weakly closed and $(E/I)_+$ is semisimple, or if J is weakly closed and $(F/J)_+$ is semisimple, then $(I \circledcirc J) \cap (E \otimes F)$ is an ideal in $E \otimes F$ (with respect to the injective cone);

(d) If I and J are weakly closed, and if at least one of $(E/I)_+$ and $(F/J)_+$ is semisimple, then $I \circledcirc J$ is an ideal in $E \circledast F$.

Proof.

(a) Since $\operatorname{lin}(\overline{E_+}^w) = {}^\perp(E'_+)$ (see §8.3), we have $\overline{\operatorname{span}(E'_+)}^{w*} = \operatorname{lin}(\overline{E_+}^w)^\perp$. Hence it follows from Lemma 10.32(b) that $(\operatorname{lin}(\overline{E_+}^w)^\perp \ltimes J) \cap (E \otimes F)$ is an ideal in $E \otimes F$. Analogously, $(I \rtimes \operatorname{lin}(\overline{F_+}^w)^\perp) \cap (E \otimes F)$ is an ideal in $E \otimes F$. The conclusions follows since the intersection of two ideals is an ideal.

(b) Analogous to (a), using Lemma 10.32(a) instead of Lemma 10.32(b).

(c) Assume that I is weakly closed and $(E/I)_+$ is semisimple (the other case is analogous). Since I is weakly closed, it follows from Lemma 10.31(b) that $I \circledcirc J = I^\perp \ltimes J$. Furthermore, by basic duality (as mentioned on page 111), the adjoint of the pushforward $E \to E/I$ is the pullback (bipositive map) $(E/I)' \cong I^\perp \to E'$,

[3] In other words, $I \cap \overline{E_+}^w$ and $J \cap \overline{F_+}^w$ are faces of $\overline{E_+}^w$ and $\overline{F_+}^w$, respectively.

so we have $(E/I)'_+ = I^\perp \cap E'_+$. Since $(E/I)_+$ is semisimple, its dual cone $(E/I)'_+$ separates points on E/I. Equivalently, the subspace $\operatorname{span}((E/I)'_+) = \operatorname{span}(I^\perp \cap E'_+)$ is weak-$*$ dense in $I^\perp$. Hence it follows from Lemma 10.32(b) that $(I^\perp \ltimes J) \cap (E \otimes F)$ is an ideal in $E \otimes F$.

(d) Analogous to (c), using Lemma 10.32(a) instead of Lemma 10.32(b). ■

Remark 10.34. In terms of the mapping properties, it is not surprising that the semisimplicity of $(E/I)_+$ and $(F/J)_+$ plays a role in Theorem 10.33. Let $\pi_I : E \to E/I$ and $\pi_J : F \to F/J$ denote the canonical maps. If both $(E/I)_+$ and $(F/J)_+$ are semisimple, then $(E/I) \circledast (F/J)$ is a proper cone (by Theorem 10.22), so now evidently $\ker(\pi_I \circledast \pi_J) = {}^\perp(I^\perp \otimes J^\perp)$ is an ideal in $E \circledast F$.

What is surprising in Theorem 10.33 is that it is sufficient for only one of $(E/I)_+$ and $(F/J)_+$ to be semisimple. This could not have been predicted solely on the basis of the mapping properties. The following example shows that we need at least one of the quotients to be semisimple, even in the finite-dimensional case.

Example 10.35. Let $E := \mathbb{R}^3$ and let $E_+ \subseteq E$ be the second-order cone $E_+ := \{(x_1, x_2, x_3) : \sqrt{x_1^2 + x_2^2} \leq x_3\}$. The injective cone $E_+^* \otimes^\varepsilon E_+$ can be identified with the cone $\mathrm{L}_+(E, E)$ of positive linear operators $E \to E$. If we identify E^* with $\mathbb{R}^3$ via the standard inner product, then E_+ is self-dual. The vectors $(1, 0, 1), (-1, 0, 1) \in \mathbb{R}^3$ define extremal rays of E_+, so the subspaces $I := \operatorname{span}\{(-1, 0, 1)\} \subseteq E^*$ and $J := \operatorname{span}\{(1, 0, 1)\} \subseteq E$ are ideals (see Proposition A.3(a)). It follows from Lemma 10.31(c) that $I \mathbin{⦶} J = I^\perp \ltimes J$. We show that this is not an ideal.

Let $b_1 \in E^* \otimes E = \mathrm{Bil}(E, E^*) \cong L(E, E)$ correspond to the identity $E \to E$, and let $b_2 \in E^* \otimes E$ be the bilinear form $E \times E^* \to \mathbb{R}$ corresponding with the linear map $(x_1, x_2, x_3) \mapsto (x_1, -x_2, x_3)$. Clearly b_1 and b_2 are positive. However, since $\dim(I^\perp) = 2$ and $\dim(J) = 1$, maps in $I^\perp \ltimes J$ cannot be invertible, so in particular we have $b_1, b_2 \notin I^\perp \ltimes J$.

It is not hard to see that $b_1 + b_2 \in I^\perp \ltimes J$, and evidently we have $0 \leq b_1 \leq b_1 + b_2$. This shows that $I^\perp \ltimes J$ is not an ideal. △

We conclude this section by providing more convenient direct formulas for the ideals $I \mathbin{⦶} J$ and $I \mathbin{⦸} J$ and their restrictions to $E \otimes F$ or $E \mathbin{\tilde{\otimes}_\alpha} F$. Roughly speaking, under certain topological assumptions we have $I \mathbin{⦶} J = (I \circledast F) + (E \circledast J)$ and $I \mathbin{⦸} J = I \circledast J$.

Ideals in the algebraic tensor product

We show that the ideals $(I \mathbin{⦸} J) \cap (E \otimes F)$ and $(I \mathbin{⦶} J) \cap (E \otimes F)$ from Theorem 10.33 are always equal to $(I \otimes J) + \operatorname{lin}(E_+ \otimes^\varepsilon F_+)$ and $(I \otimes F) + (E \otimes J)$, respectively

Lemma 10.36. *If $I \subseteq E$ and $J \subseteq F$ are subspaces, then*

$$(I^\perp \ltimes J) \cap (I \rtimes J^\perp) \cap (E \otimes F) \;=\; (I \otimes F) + (E \otimes J) + (\overline{I}^{\,w} \otimes \overline{J}^{\,w}).$$

Proof. Choose an algebraic decomposition $E \cong E_1 \oplus E_2 \oplus E_3$ with $E_1 \cong I$ and $E_1 \oplus E_2 \cong \overline{I}^{\,w}$, and likewise for $F \cong F_1 \oplus F_2 \oplus F_3$. Then $E \otimes F \cong \bigoplus_{i=1}^3 \bigoplus_{j=1}^3 (E_i \otimes F_j)$. Under this identification, $(I^\perp \ltimes J) \cap (E \otimes F)$ corresponds with those elements that are

zero in the $E_3 \otimes F_2$ and $E_3 \otimes F_3$ components. Likewise, $(I \rtimes J^\perp) \cap (E \otimes F)$ corresponds with those elements that are zero in the $E_2 \otimes F_3$ and $E_3 \otimes F_3$ components, and the conclusion follows. ∎

Corollary 10.37. *Let $I \subseteq E$ and $J \subseteq F$ be subspaces.*

(a) If at least one of I and J is weakly closed, then

$$(I^\perp \ltimes J) \cap (I \rtimes J^\perp) \cap (E \otimes F) \;=\; (I \otimes F) + (E \otimes J).$$

(b) If both I and J are weakly closed, then

$$({}^\perp(I^\perp \otimes J^\perp)) \cap (E \otimes F) \;=\; (I \otimes F) + (E \otimes J).$$

(c) The lineality space of the injective cone (in $E \otimes F$) is

$$\lin(E_+ \otimes^\varepsilon F_+) \;=\; (\lin(\overline{E_+}^{\,w}) \otimes F) + (E \otimes \lin(\overline{F_+}^{\,w})).$$

Proof.

(a) If I is weakly closed, then $\overline{I}^{\,w} \otimes \overline{J}^{\,w} \subseteq I \otimes F$, so the result follows from Lemma 10.36.

(b) Immediate, for now we have ${}^\perp(I^\perp \otimes J^\perp) = I^\perp \ltimes J = I \rtimes J^\perp$, by Lemma 10.31(c).

(c) By Proposition 10.21, we have $\lin(E_+ \otimes^\varepsilon F_+) = ({}^\perp(\lin(\overline{E_+}^{\,w})^\perp \otimes \lin(\overline{F_+}^{\,w})^\perp)) \cap (E \otimes F)$, where we note that $\lin(\overline{E_+}^{\,w})$ and $\lin(\overline{F_+}^{\,w})$ are weakly closed subspaces. ∎

Theorem 10.38. *Let $I \subseteq E$ and $J \subseteq F$ be ideals with respect to $\overline{E_+}^{\,w}$ and $\overline{F_+}^{\,w}$.*

(a) One has $(I \oslash J) \cap (E \otimes F) = (I \otimes J) + \lin(E_+ \otimes^\varepsilon F_+)$, and this is an ideal in $E \otimes F$ (with respect to the injective one);

(b) If I is weakly closed and $(E/I)_+$ is semisimple, or if J is weakly closed and $(F/J)_+$ is semisimple, then one has $(I \otimes\!\!\!\!\bigcirc J) \cap (E \otimes F) = (I \otimes F) + (E \otimes J)$, and this is an ideal in $E \otimes F$ (with respect to the injective cone).

Proof.

(a) Every ideal contains the lineality space, so we may choose a decomposition $E \cong E_1 \oplus E_2 \oplus E_3$ with $E_1 \cong \lin(\overline{E_+}^{\,w})$ and $E_1 \oplus E_2 \cong I$, and likewise for $F \cong F_1 \oplus F_2 \oplus F_3$. With respect to the decomposition $E \otimes F \cong \bigoplus_{i=1}^3 \bigoplus_{j=1}^3 (E_i \otimes F_j)$, the subspace $(\lin(\overline{E_+}^{\,w})^\perp \ltimes J) \cap (E \otimes F)$ corresponds with those elements that are zero in the $E_2 \otimes F_3$ and $E_3 \otimes F_3$ components, and $(I \rtimes \lin(\overline{F_+}^{\,w})^\perp) \cap (E \otimes F)$ corresponds with those elements that are zero in the $E_3 \otimes F_2$ and $E_3 \otimes F_3$ components. Since $\lin(E_+ \otimes^\varepsilon F_+) = (E_1 \otimes F) + (E \otimes F_1)$ (by Corollary 10.37(c)) and $I \otimes J = (E_1 \oplus E_2) \otimes (F_1 \oplus F_2)$, the conclusion follows. (This is an ideal by Theorem 10.33(a).)

(b) The formula follows from Corollary 10.37(a), and this is an ideal by Theorem 10.33(c). ∎

Ideals in the space of separately weak-$*$ continuous bilinear forms

Let $I \subseteq E$ and $J \subseteq F$ be subspaces, and write $I_+ := I \cap \overline{E_+}^w$ and $J_+ := J \cap \overline{F_+}^w$. If we let I and J belong to the dual pairs $\langle I, E'/I^\perp\rangle$, $\langle J, F'/J^\perp\rangle$, then the inclusions $T : I \hookrightarrow E$, $S : J \hookrightarrow F$ are weakly continuous (weak homomorphisms in fact; see §8.2) and approximately bipositive. Therefore $T \circledast S : I \circledast J \to E \circledast F$ is injective and bipositive, by Lemma 10.15(b).[4] In other words, we may interpret $I \circledast J$ as a subspace of $E \circledast F$, and moreover $(I \circledast J)_+ = (I \circledast J) \cap (E \circledast F)_+$.

Lemma 10.39. *The image of $I \circledast J$ under the natural inclusion $T \circledast S : I \circledast J \hookrightarrow E \circledast F$ is equal to $(E' \ltimes J) \cap (I \rtimes F')$.*

Proof. By definition (see §10.2), the map $\mathrm{id}_E \circledast S : E \circledast J \hookrightarrow E \circledast F$ is given by $((\mathrm{id}_E \circledast S)b)(\varphi, \psi) = b(\varphi, S'\psi)$. Therefore the following diagram commutes:

$$\begin{array}{ccc} E \circledast J = \mathfrak{Bil}(E'_{w*} \times J'_{w*}) & \xrightarrow{\sim} & \mathfrak{L}(E'_{w*}, J_w) \\ \downarrow{\scriptstyle \mathrm{id}_E \circledast S} & & \downarrow{\scriptstyle R \mapsto SR} \\ E \circledast F = \mathfrak{Bil}(E'_{w*} \times F'_{w*}) & \xrightarrow{\sim} & \mathfrak{L}(E'_{w*}, F_w). \end{array}$$

An operator $T \in \mathfrak{L}(E'_{w*}, F_w)$ lies in the image of $\mathfrak{L}(E'_{w*}, J_w)$ if and only if $T[E'] \subseteq J$. Therefore a bilinear form $b \in E \circledast F$ is the extension of a bilinear form in $E \circledast J$ if and only if $b \in E' \ltimes J$. By the same argument, $I \circledast J = (E \circledast J) \cap (I \rtimes F')$, and the conclusion follows. ∎

We will henceforth identify $I \circledast J$ with the subspace $(E' \ltimes J) \cap (I \rtimes F') \subseteq E \circledast F$.

Next, we turn to complementary decompositions. We say that a subspace $E_1 \subseteq E$ is *weakly complemented* if it is complemented in the weak topology. (Recall that complemented subspaces and their complements are automatically closed: if $P : E \to E$ is a continuous projection, then $\ker(P)$ and $\mathrm{ran}(P) = \ker(\mathrm{id}_E - P)$ are closed.)

Lemma 10.40. *If $E_1 \subseteq E$ and $F_1 \subseteq F$ are weakly complemented subspaces with complements $E_2 \subseteq E$ and $F_2 \subseteq F$, respectively, then $E \circledast F$ decomposes as the internal (algebraic) direct sum*

$$E \circledast F = (E_1 \circledast F_1) \oplus (E_1 \circledast F_2) \oplus (E_2 \circledast F_1) \oplus (E_2 \circledast F_2).$$

Proof. The complementary pairs give rise to weakly continuous complementary decompositions $E_w \cong (E_1)_w \times (E_2)_w$ and $F_w \cong (F_1)_w \times (F_2)_w$ (topological products). Dualizing the first of these, we obtain a weak-$*$ continuous complementary decomposition $E'_{w*} \cong (E_2^\perp)_{w*} \oplus (E_1^\perp)_{w*}$ (locally convex sum).[5] Using the mapping properties of

[4] That $T \circledast S$ is injective follows from Remark 10.18. This is a classical result; see also [Köt79, §44.4.(5)].

[5] Note that the indices are reversed when passing to the dual: we have $(E_1)'_{w*} \cong E'_{w*}/E_1^\perp \cong (E_2^\perp)_{w*}$ and vice versa.

locally convex sums and topological products (see e.g. [Köt79, §39.8]), we find

$$\begin{aligned} E \circledast F &= \mathfrak{Bil}(E'_{w*} \times F'_{w*}) \\ &\cong \mathfrak{L}(E'_{w*}, F_w) \\ &= \mathfrak{L}\left((E_2^\perp)_{w*} \oplus (E_1^\perp)_{w*}\,,\, (F_1)_w \times (F_2)_w\right) \\ &\cong \prod_{i\in\{2,1\}} \prod_{j\in\{1,2\}} \mathfrak{L}\left((E_i^\perp)_{w*}, (F_j)_w\right) \\ &\cong \prod_{i\in\{1,2\}} \prod_{j\in\{1,2\}} (E_i)_w \circledast (F_j)_w. \end{aligned}$$ ■

Corollary 10.41.

(a) If $E_1 \subseteq E$ and $F_1 \subseteq F$ are weakly complemented subspaces with complements $E_2 \subseteq E$ and $F_2 \subseteq F$, respectively, then

$$^\perp(E_1^\perp \otimes F_1^\perp) \;=\; (E_1 \circledast F_1) \oplus (E_1 \circledast F_2) \oplus (E_2 \circledast F_1),$$

where the orthogonal complement is taken with respect to the dual pair $\langle E \circledast F, E' \otimes F' \rangle$.

(b) If $\operatorname{lin}(\overline{E_+}^w)$ and $\operatorname{lin}(\overline{F_+}^w)$ are weakly complemented with complements X and Y, then

$$\operatorname{lin}((E \circledast F)_+) \;=\; \left(\operatorname{lin}(\overline{E_+}^w) \circledast \operatorname{lin}(\overline{F_+}^w)\right) \oplus \left(\operatorname{lin}(\overline{E_+}^w) \circledast Y\right) \oplus \left(X \circledast \operatorname{lin}(\overline{F_+}^w)\right).$$

We can now give concrete descriptions of the subspaces $I \circledcirc J$ and $I \circledcirc J$ from Theorem 10.33.

Theorem 10.42. *Let $I \subseteq E$ and $J \subseteq F$ be weakly closed ideals with respect to $\overline{E_+}^w$ and $\overline{F_+}^w$.*

(a) If $\operatorname{lin}(\overline{E_+}^w)$ and $\operatorname{lin}(\overline{F_+}^w)$ are weakly complemented, then $I \circledcirc J = (I \circledast J) + \operatorname{lin}((E \circledast F)_+)$, and this is an ideal in $E \circledast F$.

(b) If I and J are weakly complemented, then $I \circledcirc J = (I \circledast F) + (E \circledast J)$. This is an ideal in $E \circledast F$ if at least one of $(E/I)_+$ and $(F/J)_+$ is semisimple.

Proof.

(a) "$\supseteq$". It follows from Theorem 10.33(b) that $I \circledcirc J$ is an ideal. Since every ideal contains the lineality space, we have $\operatorname{lin}((E \circledast F)_+) \subseteq I \circledcirc J$. Furthermore, we have $E' \ltimes J \subseteq \operatorname{lin}(\overline{E_+}^w)^\perp \ltimes J$ and $I \rtimes F' \subseteq I \rtimes \operatorname{lin}(\overline{F_+}^w)^\perp$, and therefore $I \circledast J \subseteq I \circledcirc J$ (by Lemma 10.39).

"$\subseteq$". The orthogonal complement of a weakly complemented subspace is weak-$*$ complemented in the dual, so we may choose weak-$*$ continuous projections $P : E' \to \operatorname{lin}(\overline{E_+}^w)^\perp \hookrightarrow E'$ and $Q : F' \to \operatorname{lin}(\overline{F_+}^w)^\perp \hookrightarrow F'$. Let $b_1 \in I \circledcirc J$ be given, and define $b_2(\varphi, \psi) = b_1(P\varphi, Q\psi)$. Evidently b_2 is separately weak-$*$ continuous,

so $b_2 \in E \circledast F$. Furthermore, b_1 and b_2 agree on $\operatorname{lin}(\overline{E_+}^w)^\perp \times \operatorname{lin}(\overline{F_+}^w)^\perp = \overline{\operatorname{span}(E'_+)}^{w*} \times \overline{\operatorname{span}(F'_+)}^{w*}$, so b_2 is positive and $b_1 - b_2 \in \operatorname{lin}((E \circledast F)_+)$ (by Proposition 10.21). Finally, since $b_1 \in (\operatorname{lin}(\overline{E_+}^w)^\perp \ltimes J) \cap (I \rtimes \operatorname{lin}(\overline{F_+}^w)^\perp)$, we have $b_2 \in (E' \ltimes J) \cap (I \rtimes F') = I \circledast J$. Therefore, $b_1 = b_2 + (b_1 - b_2) \in (I \circledast F) + \operatorname{lin}((E \circledast F)_+)$.

(b) By Lemma 10.31(c), we have $I \circledcirc J = {}^\perp(I^\perp \otimes J^\perp)$, so the direct formula follows from Corollary 10.41(a). The conditions for $I \circledcirc J$ to be an ideal follow from Theorem 10.33(d). ■

Corollary 10.43. *If $\overline{E_+}^w$ and $\overline{F_+}^w$ are proper cones, and if $I \subseteq E$ and $J \subseteq F$ are weakly closed ideals with respect to $\overline{E_+}^w$ and $\overline{F_+}^w$, then $I \circledast J$ is an ideal in $E \circledast F$.*

Ideals in completed locally convex tensor products

Finally, we turn our attention to ideals in the completed tensor product $E \mathbin{\tilde{\otimes}_\alpha} F$. The ideals $I \circledcirc J$ and $I \circledcirc J$ obtained in Theorem 10.33 can be restricted to ideals in $E \mathbin{\tilde{\otimes}_\alpha} F$ (with respect to the injective cone). However, although we were able to find more convenient formulas for the intersections of $I \circledcirc J$ and $I \circledcirc J$ with the algebraic tensor product $E \otimes F$ (see Theorem 10.38), there are no similar formulas for the intersections with $E \mathbin{\tilde{\otimes}_\alpha} F$. To illustrate the obstruction, we first rephrase the problem in the more common terminology of normed tensor products.

Let E and F be Banach spaces, let $E_+ \subseteq E$ and $F_+ \subseteq F$ be closed proper cones, and let $J \subseteq F$ be a closed order ideal. Then $E \circledast F \cong \mathfrak{L}(E'_{w*}, F_w) \subseteq \mathfrak{L}(E', F)$ is the subspace of those operators $T : E' \to F$ for which the range of the adjoint $T' : F' \to E''$ is contained in E. By Theorem 10.33, the subspace $\{0\} \circledcirc J = E \circledcirc J = E' \ltimes J$ is an ideal in $E \circledast F$. The elements of this ideal are simply the weak-$*$-to-weak continuous operators $E' \to F$ whose range is contained in J. In particular, if α is a tensor norm, then $(E' \ltimes J) \cap (E \mathbin{\tilde{\otimes}_\alpha} F) = \mathfrak{L}(E', J) \cap (E \mathbin{\tilde{\otimes}_\alpha} F)$. It is well-known that this can be different from $E \mathbin{\tilde{\otimes}_\alpha} J$. We give two examples.

Example 10.44. If F has the approximation property but J does not, then one has $E' \mathbin{\tilde{\otimes}_\varepsilon} J \neq \mathfrak{K}(E, J)$ for some appropriate Banach space E, but also $E' \mathbin{\tilde{\otimes}_\varepsilon} F = \mathfrak{K}(E, F)$ (see [DF93, §5.3]). Therefore $\mathfrak{L}(E, J) \cap (E' \mathbin{\tilde{\otimes}_\varepsilon} F) = \mathfrak{K}(E, J)$ is strictly larger than $E' \mathbin{\tilde{\otimes}_\varepsilon} F$. △

Example 10.45. It is well-known that the operator ideal of nuclear operators is not injective: if $J \subseteq F$ is a closed subspace and if $T : E \to J$ is nuclear as a map $E \to F$, then it does not necessarily follow that T is also nuclear as a map $E \to J$ (see [DF93, §9.7]). Even if all spaces have the approximation property, so that $E' \mathbin{\tilde{\otimes}_\pi} F = \mathfrak{N}(E, F)$ and $E' \mathbin{\tilde{\otimes}_\pi} J = \mathfrak{N}(E, J)$, it can happen that $\mathfrak{L}(E, J) \cap \mathfrak{N}(E, F) \neq \mathfrak{N}(E, J)$, so that $\mathfrak{L}(E, J) \cap (E' \mathbin{\tilde{\otimes}_\pi} F) \neq E' \mathbin{\tilde{\otimes}_\pi} J$. △

The obstruction is a purely topological one, and has nothing to do with cone-theoretic issues. Therefore we only sketch the proofs of the following special cases, where a convenient formula can be obtained.

Theorem 10.46 (Injective topology; approximation property)**.** *Let E and F be complete locally convex spaces, let $E_+ \subseteq E$, $F_+ \subseteq F$ be closed proper cones, and let $I \subseteq E$, $J \subseteq F$ be closed ideals. If I or J has the approximation property, then $I \tilde{\otimes}_\varepsilon J$ is an ideal in $E \tilde{\otimes}_\varepsilon F$.*

Proof sketch. The *ε-product* $E \varepsilon F$ is the subspace of $E \circledast F$ consisting of those operators $b \in \mathfrak{Bil}(E'_{w*}, F'_{w*}) \cong \mathfrak{L}(E'_{w*}, F_w) \cong \mathfrak{L}(F'_{w*}, E_w)$ that map equicontinuous subsets of E' in relatively compact sets in F, or equivalently, that map equicontinuous subsets of F' in relatively compact sets in E (see [Köt79, §43.3.(2)]). This property is passed to subspaces, so $(E \varepsilon F) \cap (I \circledast J) \subseteq I \varepsilon J$.

Since E and F are complete, we have $E \tilde{\otimes}_\varepsilon F \subseteq E \varepsilon F$ (see [Köt79, §43.3.(5)]). Furthermore, since I and J are complete and I or J has the approximation property, we have $I \tilde{\otimes}_\varepsilon J = I \varepsilon J$ (see [Köt79, §43.3.(7)]). It follows that

$$(I \circledast J) \cap (E \tilde{\otimes}_\varepsilon F) \subseteq (I \circledast J) \cap (E \varepsilon F) \subseteq I \varepsilon J = I \tilde{\otimes}_\varepsilon J.$$

On the other hand, one clearly has $I \tilde{\otimes}_\varepsilon J \subseteq (I \circledast J) \cap (E \tilde{\otimes}_\varepsilon F)$, so we have equality. By Corollary 10.43, $I \circledast J$ is an ideal in $E \circledast F$, so it follows that $I \tilde{\otimes}_\varepsilon J$ is an ideal in $E \tilde{\otimes}_\varepsilon F$. ■

If E and F are Banach spaces, then the ε-product in the proof of Theorem 10.46 can be replaced by a suitable space of compact operators.

The second situation where a more convenient formula can be obtained is if the subspaces are complemented. Let us say that a *locally convex tensor topology* is a locally convex topology α defined on $E \otimes F$ for every pair (E, F) of locally convex spaces such that:

(i) α is a compatible topology on $E \otimes F$ for every pair (E, F);

(ii) α satisfies the *continuous mapping property*: if $T : E \to G$ and $S : F \to H$ are continuous, then $T \otimes S : E \otimes_\alpha F \to G \otimes_\alpha H$ is also continuous.

Examples of locally convex tensor topologies include the projective topology π and the injective topology ε. More generally, every tensor norm gives rise to a locally convex tensor topology that even satisfies the equicontinuous mapping property; see [DF93, §35.2].

Theorem 10.47 (Arbitrary topology; complemented subspaces)**.** *Let E and F be complete locally convex spaces, let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones, let α be a locally convex tensor topology, and let $I \subseteq E$, $J \subseteq F$ be closed ideals with respect to $\overline{E_+}$ and $\overline{F_+}$.*

(a) If I, J, $\operatorname{lin}(\overline{E_+})$ and $\operatorname{lin}(\overline{F_+})$ are complemented, then $(I \mathbin{⍁} J) \cap (E \tilde{\otimes}_\alpha F) = (I \tilde{\otimes}_\alpha F) + \operatorname{lin}(E_+ \tilde{\otimes}^\varepsilon_\alpha F_+)$, and this is an ideal in $E \tilde{\otimes}_\alpha F$ (with respect to the injective cone).

(b) If I and J are complemented, then $(I \mathbin{⍀} J) \cap (E \tilde{\otimes}_\alpha F) = (I \tilde{\otimes}_\alpha F) + (E \tilde{\otimes}_\alpha J)$. This is an ideal in $E \tilde{\otimes}_\alpha F$ (with respect to the injective cone) if at least one of $(E/I)_+$ and $(F/J)_+$ is semisimple.

Proof sketch. Given closed subspaces $E_1, \dots, E_n \subseteq E$, we say that $E \cong \bigoplus_{i=1}^n E_i$ *topologically* if the canonical map $\bigoplus_{i=1}^n E_i \to E$ is a topological isomorphism. Equivalently, if $E \cong \bigoplus_{i=1}^n E_i$ algebraically, then one has $E \cong \bigoplus_{i=1}^n E_i$ topologically if and only if every projection $E \to E_i$ is continuous (see [Sch99, Theorem 2.2]).

If $E \cong \bigoplus_{i=1}^n E_i$ topologically and $F \cong \bigoplus_{j=1}^m F_j$ topologically, then $E \mathbin{\tilde{\otimes}_\alpha} F \cong \bigoplus_{i,j}(E_i \mathbin{\tilde{\otimes}_\alpha} F_j)$ topologically. Analogously, Lemma 10.40 can be extended to prove that $E \circledast F \cong \bigoplus_{i,j}(E_i \circledast F_j)$, and the following diagram commutes:

$$\begin{array}{ccccc} E \mathbin{\tilde{\otimes}_\alpha} F & \longrightarrow & E \mathbin{\tilde{\otimes}_\varepsilon} F & \longrightarrow & E \circledast F \\ \downarrow \wr & & \downarrow \wr & & \downarrow \wr \\ \bigoplus_{i,j}(E_i \mathbin{\tilde{\otimes}_\alpha} F_j) & \longrightarrow & \bigoplus_{i,j}(E_i \mathbin{\tilde{\otimes}_\varepsilon} F_j) & \longrightarrow & \bigoplus_{i,j}(E_i \circledast F_j). \end{array}$$

In particular, for every subset $A \subseteq [n] \times [m]$ we have

$$\left(\bigoplus_{(i,j)\in A} (E_i \circledast F_j) \right) \cap (E \mathbin{\tilde{\otimes}_\alpha} F) = \bigoplus_{(i,j)\in A} (E_i \mathbin{\tilde{\otimes}_\alpha} F_j).$$

Therefore the result follows from Theorem 10.42, using the following decompositions:

(a) $E \cong E_1 \oplus E_2 \oplus E_3$ and $F \cong F_1 \oplus F_2 \oplus F_3$ (topologically), where $E_1 = \operatorname{lin}(\overline{E_+})$, $E_1 \oplus E_2 = I$, $F_1 = \operatorname{lin}(\overline{F_+})$, $F_1 \oplus F_2 = J$; and $A = \{(1,1), (1,2), (1,3), (2,1), (2,2), (3,1)\}$.

(b) $E \cong E_1 \oplus E_2$ and $F \cong F_1 \oplus F_2$, where $E_1 = I$, $F_1 = J$; and $A = \{(1,1), (1,2), (2,1)\}$. ■

10.6 Extremal rays of the injective cone

As an application of the results from §10.5, we show that the tensor product of two extremal rays defines an extremal ray of the injective cone. In §11.1 we will prove that all rank one extremal directions of the injective cone are of this form, but Example 10.51 will show that there might be extremal directions of larger rank.

Proposition 10.48. *If $x_0 \in \overline{E_+}^w \setminus \{0\}$ and $y_0 \in \overline{F_+}^w \setminus \{0\}$ define extremal rays of $\overline{E_+}^w$ and $\overline{F_+}^w$, then $x_0 \otimes y_0 \in E \otimes F \subseteq E \circledast F$ defines an extremal ray of $(E \circledast F)_+$. In other words:*

$$\operatorname{rext}(\overline{E_+}^w) \otimes_s \operatorname{rext}(\overline{F_+}^w) \subseteq \operatorname{rext}((E \circledast F)_+).$$

Proof. Let $M := \{\lambda x_0 : \lambda \geq 0\}$ denote the ray generated by x_0. Then M is an extremal ray, so in particular a face. Every face contains the minimal face $\operatorname{lin}(\overline{E_+}^w)$, but M does not contain a non-trivial subspace, so $\overline{E_+}^w$ is a proper cone. Furthermore, $I := \operatorname{span}(M) = \operatorname{span}(x_0)$ is an ideal by Proposition A.3(a), and is weakly closed because it is finite-dimensional. Analogously, $J := \operatorname{span}(y_0)$ is a weakly closed ideal in F, so it follows from Corollary 10.43 that $I \otimes J$ defines an ideal in $E \circledast F$. To complete the proof, note that $x_0 \otimes y_0 \in (E \circledast F)_+$, and that $-x_0 \otimes y_0 \notin (E \circledast F)_+$ because $(E \circledast F)_+$ is a proper cone. In other words, $(I \otimes J)_+$ is the ray generated by $x_0 \otimes y_0$. ■

Corollary 10.49. *If $\langle E, E'\rangle$, $\langle F, F'\rangle$ are dual pairs and if $E_+ \subseteq E$, $F_+ \subseteq F$ are convex cones, then*

$$\operatorname{rext}(\overline{E_+}^{\,w}) \otimes_s \operatorname{rext}(\overline{F_+}^{\,w}) \subseteq \operatorname{rext}(E_+ \otimes^{\varepsilon} F_+).$$

Corollary 10.50. *If E, F are complete locally convex spaces, if $E_+ \subseteq E$, $F_+ \subseteq F$ are convex cones, and if α is a compatible locally convex topology on $E \otimes F$ for which the natural map $E \tilde{\otimes}_\alpha F \to E \tilde{\otimes}_\varepsilon F$ is injective, then*

$$\operatorname{rext}(\overline{E_+}^{\,w}) \otimes_s \operatorname{rext}(\overline{F_+}^{\,w}) \subseteq \operatorname{rext}(E_+ \tilde{\otimes}^{\varepsilon}_{\alpha} F_+).$$

Note: if $E \tilde{\otimes}_\alpha F \to E \tilde{\otimes}_\varepsilon F$ is not injective, then $E_+ \tilde{\otimes}^{\varepsilon}_{\alpha} F_+$ does not have extremal rays, since it is not a proper cone (see Corollary 10.23).

In Theorem 9.22, we found that the extremal rays of the projective cone are precisely the tensor products of the extremal rays of the base cones. This is not true for the injective cone; the following example shows that the inclusion from Corollary 10.49 can be strict.

Example 10.51 (cf. Example 9.7, Example 10.20)**.** Let E be finite-dimensional, and let $E_+ \subseteq E$ be a proper, generating, polyhedral cone which is *not* a simplex cone. Then both $E_+^* \otimes^{\pi} E_+$ and $E_+^* \otimes^{\varepsilon} E_+$ are proper, generating, polyhedral cones (use that the class of proper, generating, polyhedral cones is closed under taking duals and projective tensor products). As such, they are generated by their extremal rays. However, it follows from [BL75, Proposition 3.1] (see also Theorem 13.2 below) that $E_+^* \otimes^{\pi} E_+ \neq E_+^* \otimes^{\varepsilon} E_+$, so in particular $\operatorname{rext}(E_+^* \otimes^{\varepsilon} E_+) \neq \operatorname{rext}(E_+^* \otimes^{\pi} E_+) = \operatorname{rext}(E_+^*) \otimes_s \operatorname{rext}(E_+)$.

It will follow from Corollary 11.4(b) below that the additional extremal directions of the injective cone must necessarily have rank ≥ 2. △

Remark 10.52. It is somewhat remarkable that the injective cone preserves extremal rays, because the injective norm does not preserve extreme points (of the closed unit ball). Indeed, if $E = F = \mathbb{R}^2$ with the Euclidean norm, then $E \otimes_\varepsilon F \cong \mathbb{R}^{2\times 2}$ with the operator norm (i.e. the Schatten ∞-norm). But the extreme points of the unit ball for the operator norm are the orthogonal matrices, which in particular have full rank. In other words, no rank 1 operator is an extreme point, so in this case $\operatorname{ext}(B_{E\otimes_\varepsilon F})$ is *disjoint* from $\operatorname{ext}(B_E) \otimes_s \operatorname{ext}(B_F)$.

This discrepancy can be explained as follows. In §9.6, we proved that the projective unit ball preserves extreme points (at least in the finite-dimensional case). The proof used homogenization: given finite-dimensional normed spaces E and F, we considered the respective "ice cream cones" in $E \oplus \mathbb{R}$ and $F \oplus \mathbb{R}$. However, the tensor product $(E \oplus \mathbb{R}) \otimes (F \oplus \mathbb{R}) \cong (E \otimes F) \oplus E \oplus F \oplus \mathbb{R}$ is larger than $(E \otimes F) \oplus \mathbb{R}$, so the projective cone is larger than the homogenization of the projective unit ball. In order to recover an extreme point of the projective unit ball, we had to work with a two-dimensional face of the projective cone. Thus, extremal rays of the projective cone do not correspond directly with extreme points of the projective unit ball. Apparently, the two-dimensional face used in this argument has no analogue in the injective cone.

Chapter 11

Reasonable cross-cones

In this chapter, we give three applications of the results from Chapter 9 and Chapter 10 to other 'reasonable' cones in the tensor product.

This chapter is based on Chapter 5 of [Dob20b].

Introduction

Apart from the projective and injective cone, there are many other reasonable cones in the tensor product. Depending on the application, other choices may be appropriate as well. For instance, the tensor product of two spaces of hermitian matrices is again a space of hermitian matrices, but if the spaces are equipped with the positive semidefinite cone, then neither the projective nor the injective cone is equal to the positive semidefinite cone in the tensor product (see e.g. [And04, §2]).

In this chapter, mirroring an analogous definition in the normed theory, we study the broader class of 'reasonable crosscones' in the tensor product.

Definition 11.1. Let $\langle E, E'\rangle$ and $\langle F, F'\rangle$ be dual pairs, and let $E_+ \subseteq E$ and $F_+ \subseteq F$ be convex cones. We say that a convex cone $\mathcal{K} \subseteq E \otimes F$ is a *reasonable cross-cone* if it satisfies the following criteria:

(i) For all $x \in E_+$ and $y \in F_+$ one has $x \otimes y \in \mathcal{K}$;

(ii) For all $\varphi \in E'_+$ and $\psi \in F'_+$, one has $\varphi \otimes \psi \in \mathcal{K}'$.

Here $\mathcal{K}'$ denotes the dual cone of $\mathcal{K}$ with respect to any reasonable dual G of $E \otimes F$ (that is, $E' \otimes F' \subseteq G \subseteq \mathfrak{Bil}(E \times F)$; see §8.2). The definition does not depend on the choice of reasonable dual, because $\varphi \otimes \psi \in E' \otimes F'$.

Reasonable cross-cones in the completed tensor product $E \mathbin{\tilde{\otimes}_\alpha} F$ (E and F locally convex, α a compatible locally convex topology on $E \otimes F$) are defined analogously.

The following proposition shows that a cone in $E \otimes F$ is a reasonable crosscone if and only if it lies somewhere between the projective and injective cones.

Proposition 11.2. *Let $\langle E, E'\rangle$, $\langle F, F'\rangle$ be dual pairs and let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones. Then $E_+ \otimes^\pi F_+$ and $E_+ \otimes^\varepsilon F_+$ are reasonable cross-cones, and $E_+ \otimes^\pi F_+ \subseteq E_+ \otimes^\varepsilon F_+$. Furthermore, a convex cone $\mathcal{K} \subseteq E \otimes F$ is a reasonable cross-cone if and only if $E_+ \otimes^\pi F_+ \subseteq \mathcal{K} \subseteq E_+ \otimes^\varepsilon F_+$.*

Proof. For $x \in E_+$, $y \in F_+$, $\varphi \in E'_+$, $\psi \in F'_+$ we have $\langle x, \varphi\rangle \geq 0$ and $\langle y, \psi\rangle \geq 0$, and therefore $\langle x \otimes y, \varphi \otimes \psi\rangle = \langle x, \varphi\rangle \cdot \langle y, \psi\rangle \geq 0$. It follows that $E_+ \otimes^\pi F_+$ and $E_+ \otimes^\varepsilon F_+$ are reasonable cross-cones and that $E_+ \otimes^\pi F_+ \subseteq E_+ \otimes^\varepsilon F_+$.

For a general convex cone $\mathcal{K} \subseteq E \otimes F$, clearly Definition 11.1(i) is equivalent to $E_+ \otimes^\pi F_+ \subseteq \mathcal{K}$, and Definition 11.1(ii) is equivalent to $\mathcal{K} \subseteq E_+ \otimes^\varepsilon F_+$. ■

Likewise, a convex cone $\mathcal{K} \subseteq E \,\tilde{\otimes}_\alpha\, F$ is a reasonable cross-cone if and only if $E_+ \,\tilde{\otimes}^\pi_\alpha\, F_+ \subseteq \mathcal{K} \subseteq E_+ \,\tilde{\otimes}^\varepsilon_\alpha\, F_+$. If this is the case, then in particular $\mathcal{K} \cap (E \otimes F)$ is a reasonable cross-cone in the algebraic tensor product $E \otimes F$.

In this chapter, we give three applications of the results from the previous chapters to arbitrary reasonable crosscones. We show that all reasonable crosscones have the same rank one tensors whenever E_+ and F_+ are weakly closed and proper (§11.1), we show that ideals and extremal rays are preserved by reasonable crosscones (§11.2), and we show that every reasonable crosscone in $E \otimes F$ is semisimple with respect to any reasonable dual space if E_+ and F_+ are semisimple (§11.3). Finally, in §11.4, we study the related problem of semisimplicity in completed locally convex tensor products, but there things are a bit more complicated.

11.1 Rank one tensors of reasonable crosscones

The definition of reasonable crosscones is based on two criteria regarding rank one tensors in $E \otimes F$ and $E' \otimes F'$. We show that, if E_+ and F_+ are sufficiently nice, then all reasonable crosscones contain the same rank one tensors (Corollary 11.4). Using this, we will classify all rank one tensors in the projective and injective cones (Proposition 11.6).

If $\langle E, E'\rangle$ is a dual pair, then we say that a convex cone $E_+ \subseteq E$ is *approximately generating* (or *total*) if $\operatorname{span}(E_+)$ is weakly dense in E.

If $\mathcal{K} \subseteq E \otimes F$ is a convex cone, then we understand $\mathcal{K}'$ to be the dual cone with respect to some reasonable dual G of $E \otimes F$. The choice of G does not matter, for we will restrict our attention to $\mathcal{K}' \cap (E' \otimes F')$.

The following result is an extension of [Bar76, Theorem 3.3].

Proposition 11.3. *Let $\langle E, E'\rangle$, $\langle F, F'\rangle$ be dual systems, let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones, and let $\mathcal{K} \subseteq E \otimes F$ be a reasonable crosscone.*

(a) If E_+ and F_+ are weakly closed proper cones, then a rank one tensor $x_0 \otimes y_0 \in E \otimes F$ belongs to $\mathcal{K}$ if and only if either $x_0 \in E_+$ and $y_0 \in F_+$ or $-x_0 \in E_+$ and $-y_0 \in F_+$.

(b) If E_+ and F_+ are approximately generating, then a rank one tensor $\varphi_0 \otimes \psi_0 \in E' \otimes F'$ defines a positive linear functional on $\mathcal{K}$ if and only if either $\varphi_0 \in E'_+$ and $\psi_0 \in F'_+$ or $-\varphi_0 \in E'_+$ and $-\psi_0 \in F'_+$.

Proof.

(a) "$\Longleftarrow$". If $x_0 \in E_+$ and $y_0 \in F_+$, then $x_0 \otimes y_0 \in \mathcal{K}$ by definition. If $-x_0 \in E_+$ and $-y_0 \in F_+$, note that $x_0 \otimes y_0 = -x_0 \otimes -y_0 \in \mathcal{K}$.

"$\Longrightarrow$". Let $x_0 \otimes y_0 \in \mathcal{K}$ be of rank one (i.e. with $x_0, y_0 \neq 0$). Weakly closed proper cones are semisimple, so the dual cones E'_+ and F'_+ separate points on E_+ and F_+, respectively. Choose $\varphi_0 \in E'_+$, $\psi_0 \in F'_+$ such that $\langle x_0, \varphi_0\rangle \neq 0$ and $\langle y_0, \psi_0\rangle \neq 0$. Then $\varphi_0 \otimes \psi_0$ defines a positive linear functional on $\mathcal{K}$, so we have $\langle x_0, \varphi_0\rangle\langle y_0, \psi_0\rangle = \langle x_0 \otimes y_0, \varphi_0 \otimes \psi_0\rangle \geq 0$. It follows that $\langle x_0, \varphi_0\rangle$ and $\langle y_0, \psi_0\rangle$ have the same sign. Since $-x_0 \otimes -y_0 = x_0 \otimes y_0$, we may assume without loss of generality that $\langle x_0, \varphi_0\rangle, \langle y_0, \psi_0\rangle > 0$.

If $\varphi \in E'_+$ is arbitrary, then $\varphi \otimes \psi_0$ is a positive linear functional on $\mathcal{K}$, so we have $\langle x_0, \varphi\rangle\langle y_0, \psi_0\rangle = \langle x_0 \otimes y_0, \varphi \otimes \psi_0\rangle \geq 0$. Since $\langle y_0, \psi_0\rangle > 0$, it follows that $\langle x_0, \varphi\rangle \geq 0$ for all $\varphi \in E'_+$, which shows that $x_0 \in (E'_+)' = \overline{E_+}^w = E_+$. Analogously, we find $y_0 \in F_+$.

(b) In this case E_+ and F_+ separate points on E' and F', so the same proof can be carried out. (If $\varphi_0 \otimes \psi_0 \in \mathcal{K}'$ has rank one, then we may choose $x_0 \in E_+$, $y_0 \in F_+$ such that $\langle x_0, \varphi_0\rangle\langle y_0, \psi_0\rangle > 0$, and use these to show that $\varphi_0 \in E'_+$ and $\psi_0 \in F'_+$ or $-\varphi_0 \in E'_+$ and $-\psi_0 \in F'_+$.) ■

Corollary 11.4. *Let $\langle E, E'\rangle$, $\langle F, F'\rangle$ be dual systems, and let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones.*

(a) If E_+ and F_+ are weakly closed proper cones, then all reasonable crosscones in $E \otimes F$ agree on the rank one tensors.

(b) The set of rank one extremal directions of the injective cone $E_+ \otimes^\varepsilon F_+ \subseteq E \otimes F$ is given by $\operatorname{rext}(\overline{E_+}^w) \otimes_s \operatorname{rext}(\overline{F_+}^w)$.

(c) If E_+ and F_+ are weakly closed proper cones, and if $\mathcal{K} \subseteq E \otimes F$ is a reasonable crosscone, then the set of rank one extremal directions of $\mathcal{K}$ is given by $\operatorname{rext}(\mathcal{K}) = \operatorname{rext}(E_+) \otimes_s \operatorname{rext}(F_+)$.

Proof.

(a) Immediate from Proposition 11.3(a).

(b) If $x_0 \in \overline{E_+}^w$ and $y_0 \in \overline{F_+}^w$ are extremal directions, then $x_0 \otimes y_0$ is an extremal direction of $E_+ \otimes^\varepsilon F_+$, by Corollary 10.49. For the converse, suppose that $x_0 \otimes y_0$ is a rank one extremal direction of $E_+ \otimes^\varepsilon F_+$. Then $E \otimes F \neq \{0\}$ (since there exist rank one tensors), so $E \neq \{0\}$ and $F \neq \{0\}$. Similarly, $E_+ \otimes^\varepsilon F_+$ is a proper cone (since it has extremal directions), so now it follows from Theorem 10.22 that $\overline{E_+}^w$ and $\overline{F_+}^w$ are proper cones. Since $E_+ \otimes^\varepsilon F_+ = \overline{E_+}^w \otimes^\varepsilon \overline{F_+}^w$, it follows from (a) that $x_0 \otimes y_0 \in \overline{E_+}^w \otimes^\pi \overline{F_+}^w$. Clearly $x_0 \otimes y_0$ is automatically extremal in this (smaller) cone, so it follows from Theorem 9.22 that x_0 and y_0 (or $-x_0$ and $-y_0$) are extremal directions of $\overline{E_+}^w$ and $\overline{F_+}^w$.

(c) By (a), every rank one extremal direction of a reasonable crosscone is also an extremal direction of every smaller reasonable crosscone. By (b) and Theorem 9.22, the projective and injective cones have the same rank one extremal directions. ■

Remark 11.5. In general $E_+ \otimes^\pi F_+$ and $E_+ \otimes^\varepsilon F_+$ do not agree on the rank one tensors. For example, if E_+ is not weakly closed, then all non-zero tensors in $E \otimes \mathbb{R} \cong E$ have rank one, but $E_+ \otimes^\pi \mathbb{R}_{\geq 0} = E_+$ whereas $E_+ \otimes^\varepsilon \mathbb{R}_{\geq 0} = \overline{E_+}^w$. As a more extreme example, let $E_+ = E$ and $F_+ = \{0\}$; then $E_+ \otimes^\pi F_+ = \{0\}$, whereas $E_+ \otimes^\varepsilon F_+ = E \otimes F$.

Using Proposition 11.3, we can determine exactly which rank one tensors belong to the projective and injective cones (without additional assumptions on E_+ and F_+).

Proposition 11.6. *Let $\langle E, E' \rangle$, $\langle F, F' \rangle$ be dual pairs and let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones.*

(a) A rank one tensor $x_0 \otimes y_0 \in E \otimes F$ belongs to the projective cone $E_+ \otimes^\pi F_+$ if and only if at least one of the following applies:

(i) $x_0 \in \operatorname{lin}(E_+)$ and $y_0 \in \operatorname{span}(F_+)$;
(ii) $x_0 \in \operatorname{span}(E_+)$ and $y_0 \in \operatorname{lin}(F_+)$;
(iii) $x_0 \in E_+$ and $y_0 \in F_+$;
(iv) $-x_0 \in E_+$ and $-y_0 \in F_+$.

(b) A rank one tensor $x_0 \otimes y_0 \in E \otimes F$ belongs to the injective cone $E_+ \otimes^\varepsilon F_+$ if and only if at least one of the following applies:

(i) $x_0 \in \operatorname{lin}(\overline{E_+}^w)$;
(ii) $y_0 \in \operatorname{lin}(\overline{F_+}^w)$;
(iii) $x_0 \in \overline{E_+}^w$ and $y_0 \in \overline{F_+}^w$;
(iv) $-x_0 \in \overline{F_+}^w$ and $-y_0 \in \overline{F_+}^w$.

Proof.

(a) "$\Longleftarrow$". If $x_0 \in E_+$ and $y_0 \in F_+$, then clearly $x_0 \otimes y_0 \in E_+ \otimes^\pi F_+$. If $x_0 \in \operatorname{lin}(E_+)$ and $y_0 \in \operatorname{span}(F_+)$, then it follows from Corollary 9.17 that $x_0 \otimes y_0 \in \operatorname{lin}(E_+ \otimes^\pi F_+) \subseteq E_+ \otimes^\pi F_+$. The other two cases are analogous.

"$\Longrightarrow$". Let $x_0 \otimes y_0 \in E_+ \otimes^\pi F_+$ be of rank one (i.e. with $x_0, y_0 \neq 0$), and write $x_0 \otimes y_0 = \sum_{i=1}^k x_i \otimes y_i$ with $x_1, \ldots, x_k \in E_+$, $y_1, \ldots, y_k \in F_+$. Note that we must have $y_0 \in \operatorname{span}(F_+)$: choose $\varphi \in E'$ such that $\varphi(x_0) = 1$, then

$$y_0 = (\varphi \otimes \mathrm{id}_F)(x_0 \otimes y_0) = (\varphi \otimes \mathrm{id}_F) \left(\sum_{i=1}^k x_i \otimes y_i \right) = \sum_{i=1}^k \varphi(x_i) y_i \in \operatorname{span}(F_+).$$

Analogously, $x_0 \in \operatorname{span}(E_+)$.

Let $\pi_{\operatorname{lin}(E_+)} : E \to E/\operatorname{lin}(E_+)$ and $\pi_{\operatorname{lin}(F_+)} : F \to F/\operatorname{lin}(F_+)$ be the canonical maps. Since $\operatorname{lin}(E_+)$ and $\operatorname{lin}(F_+)$ are ideals, the quotient cones are proper (see Appendix A.1). For notational convenience, let $x_0', \ldots, x_k'$ and $y_0', \ldots, y_k'$ denote the images of $x_0, \ldots, x_k$ and $y_0, \ldots, y_k$ in the respective quotients. Now $x_0' \otimes y_0' \in (E/\operatorname{lin}(E_+))_+ \otimes^\pi (F/\operatorname{lin}(F_+))_+$ has rank at most one. If it has rank

zero, then $x_0 \in \lin(E_+)$ or $y_0 \in \lin(F_+)$, so we are done. Assume therefore that $x_0' \otimes y_0'$ has rank one.

Define $X := \operatorname{span}\{x_0', \dots, x_k'\} \subseteq E/\lin(E_+)$, and let $X_+ \subseteq X \cap (E/\lin(E_+))_+$ be the convex cone generated by $x_1', \dots, x_k'$. Then X is finite-dimensional and X_+ is closed (because it is finitely generated) and proper (since it is contained in the proper cone $(E/\lin(E_+))_+$). Define $Y_+ \subseteq F/\lin(F_+)$ and $Y \subseteq F/\lin(F_+)$ analogously.

Since $x_0', \dots, x_k'$ and $y_0', \dots, y_k'$ belong to X and Y, it follows that $x_0' \otimes y_0' = \sum_{i=1}^k x_i' \otimes y_i'$ holds in $X \otimes Y$, so we have $x_0' \otimes y_0' \in X_+ \otimes^\pi Y_+$. Since X_+ and Y_+ are closed and proper, it follows from Proposition 11.3(a) that $x_0' \in X_+$ and $y_0' \in Y_+$ or $-x_0' \in X_+$ and $-y_0' \in Y_+$. Since the quotient maps $\pi_{\lin(E_+)}$ and $\pi_{\lin(F_+)}$ are bipositive (see Proposition A.6), it follows that $x_0 \in E_+$ and $y_0 \in F_+$ or $-x_0 \in E_+$ and $-y_0 \in F_+$.

(b) "$\Longleftarrow$". If $x_0 \in \overline{E_+}^w$ and $y_0 \in \overline{F_+}^w$, then

$$x_0 \otimes y_0 \in \overline{E_+}^w \otimes^\pi \overline{F_+}^w \subseteq \overline{E_+}^w \otimes^\varepsilon \overline{F_+}^w = E_+ \otimes^\varepsilon F_+.$$

If $x_0 \in \lin(\overline{E_+}^w)$, $y_0 \in F$, then $x_0 \otimes y_0 \in \lin(E_+ \otimes^\varepsilon F_+) \subseteq E_+ \otimes^\varepsilon F_+$, by Corollary 10.37(c). The other two cases are analogous.

"$\Longrightarrow$". Let $x_0 \otimes y_0 \in E_+ \otimes^\varepsilon F_+$ be of rank one (i.e. with $x_0, y_0 \neq 0$). If $x_0 \in \lin(\overline{E_+}^w)$ or $y_0 \in \lin(\overline{F_+}^w)$, then we are done, so assume $x_0 \notin \lin(\overline{E_+}^w)$ and $y_0 \notin \lin(\overline{F_+}^w)$. Since $\lin(\overline{E_+}^w) = {}^\perp(E_+')$, this means that we may choose $\varphi_0 \in E_+'$, $\psi_0 \in F_+'$ such that $\langle x_0, \varphi_0 \rangle \neq 0$ and $\langle y_0, \psi_0 \rangle \neq 0$. Now it follows from the argument of Proposition 11.3 that either $x_0 \in \overline{E_+}^w$ and $y_0 \in \overline{F_+}^w$, or $-x_0 \in \overline{E_+}^w$ and $-y_0 \in \overline{F_+}^w$. ■

Proposition 11.6 can be paraphrased as follows: every rank one tensor in the projective or injective cone is either positive for obvious reasons (conditions *(iii)* and *(iv)*) or belongs to the lineality space (conditions *(i)* and *(ii)*).

11.2 Ideals and faces of reasonable crosscones

An ideal with respect to the injective cone is also an ideal with respect to every smaller cone, and a face of the injective cone is also a face of every smaller cone containing that face. Therefore the results from Chapter 10 immediately give rise to the following consequences (among others).

Proposition 11.7. *Let $\langle E, E' \rangle$, $\langle F, F' \rangle$ be dual pairs, let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones, and let $\mathcal{K} \subseteq E \otimes F$ be a reasonable crosscone. Then:*

(a) If E_+ and F_+ are weakly closed and if $I \subseteq E$, $J \subseteq F$ are ideals, then $(I \otimes J) + \lin(E_+ \otimes^\varepsilon F_+)$ is an ideal with respect to $\mathcal{K}$. Additionally, if I is weakly closed and $(E/I)_+$ is semisimple, or if J is weakly closed and $(F/J)_+$ is semisimple, then $(I \otimes F) + (E \otimes J)$ is an ideal with respect to $\mathcal{K}$.

(b) The lineality space of $\mathcal{K}$ satisfies

$$(\operatorname{lin}(E_+) \otimes \operatorname{span}(F_+)) + (\operatorname{span}(E_+) \otimes \operatorname{lin}(F_+)) \subseteq \operatorname{lin}(\mathcal{K}) \subseteq (\operatorname{lin}(\overline{E_+}^w) \otimes F) + (E \otimes \operatorname{lin}(\overline{F_+}^w)).$$

(c) If E_+ and F_+ are weakly closed and if $x_0 \in E_+$, $y_0 \in F_+$ define extremal rays, then $x_0 \otimes y_0$ defines an extremal ray of $\mathcal{K}$.

11.3 Semisimplicity of reasonable crosscones in the algebraic tensor product

Recall that a convex cone E_+ is *semisimple* if it is contained in a weakly closed proper cone, or equivalently, if E'_+ separates points on E. In this section, we prove that every reasonable crosscone in $E \otimes F$ is semisimple if E_+ and F_+ are semisimple, and we determine necessary and sufficient criteria for the projective and injective cones to be semisimple. Similar results in completed locally convex tensor products will be discussed in §11.4.

We start by setting up a partial converse, using the following proposition.

Proposition 11.8. *Let $\langle E, E'\rangle$, $\langle F, F'\rangle$ be dual pairs, and let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones. If G is a reasonable dual of $E \otimes F$ and if $\mathcal{K} \subseteq E \otimes F$ is a reasonable crosscone, then*

$$\overline{E_+}^w \otimes^\pi \overline{F_+}^w \subseteq \overline{\mathcal{K}}^w,$$

where $\overline{\mathcal{K}}^w$ denotes the $\sigma(E \otimes F, G)$-closure of $\mathcal{K}$.

Proof. Let $\mathfrak{T}_i$ denote the finest compatible topology on the tensor product $E_w \otimes F_w$ (known as the *inductive* topology, not to be confused with the injective topology). Then the natural map $E \times F \to E \otimes F$ is separately continuous as a map $E_w \times F_w \to (E \otimes F, \mathfrak{T}_i)$, and the dual of $(E \otimes F, \mathfrak{T}_i)$ is $\mathfrak{Bil}(E_w \times F_w) = \mathfrak{Bil}(E \times F)$; see [Köt79, §44.1.(5)]. In particular, since $G \subseteq \mathfrak{Bil}(E \times F)$, it follows that $w = \sigma(E \otimes F, G)$ is weaker than $\mathfrak{T}_i$. Therefore: $\overline{\mathcal{K}}^i \subseteq \overline{\mathcal{K}}^w$.

To complete the proof, we show that $\overline{E_+}^w \otimes^\pi \overline{F_+}^w \subseteq \overline{\mathcal{K}}^i$. Since $\mathcal{K}$ is a reasonable crosscone, we have $E_+ \otimes^\pi F_+ \subseteq \mathcal{K} \subseteq \overline{\mathcal{K}}^i$. Since $E_w \times F_w \to (E \otimes F, \mathfrak{T}_i)$ is separately continuous, for every $x_0 \in E_+$ one has $x_0 \otimes \overline{F_+}^w \subseteq \overline{\mathcal{K}}^i$. (The inverse image of $\overline{\mathcal{K}}^i$ under the map $y \mapsto x_0 \otimes y$ contains F_+, and therefore $\overline{F_+}^w$.) Then, by the same argument, for every $y_0 \in \overline{F_+}^w$ we have $\overline{E_+}^w \otimes y_0 \subseteq \overline{\mathcal{K}}^i$. It follows that $\overline{E_+}^w \otimes_s \overline{F_+}^w \subseteq \overline{\mathcal{K}}^i$, and the result follows by taking positive combinations. ∎

For clarity, let us say that $\mathcal{K}$ is *G-semisimple* if it is semisimple for the dual pair $\langle E \otimes F, G\rangle$ (i.e. if $\overline{\mathcal{K}}^{\sigma(E\otimes F, G)}$ is a proper cone), where G is a reasonable dual of $E \otimes F$.

Theorem 11.9. *Let $\langle E, E'\rangle$, $\langle F, F'\rangle$ be dual pairs, let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones, let G be a reasonable dual of $E \otimes F$, and let $\mathcal{K} \subseteq E \otimes F$ be a reasonable crosscone.*

(a) If E_+ and F_+ are semisimple, then $\mathcal{K}$ is G-semisimple.

(b) If $E_+ \neq \{0\}$ and $F_+ \neq \{0\}$, and if $\mathcal{K}$ is G-semisimple, then E_+ and F_+ are semisimple.

Proof.

(a) Semisimplicity means that $\overline{E_+}^w$ and $\overline{F_+}^w$ are proper cones, so it follows from Theorem 10.22 that $E_+ \otimes^\varepsilon F_+$ is a proper cone. Furthermore, $E_+ \otimes^\varepsilon F_+$ is weakly closed (see Remark 10.1), so it follows that $\mathcal{K}$ is contained in a weakly closed proper cone.

(b) It follows from Proposition 11.8 that $\overline{E_+}^w \otimes^\pi \overline{F_+}^w \subseteq \overline{\mathcal{K}}^w$, where $\overline{\mathcal{K}}^w$ is a proper cone (by semisimplicity). In particular, $\overline{E_+}^w \otimes^\pi \overline{F_+}^w$ is a proper cone. By assumption, we have $\overline{E_+}^w, \overline{F_+}^w \neq \{0\}$, so it follows from Theorem 9.10 that $\overline{E_+}^w$ and $\overline{F_+}^w$ must be proper cones as well. Equivalenty: E_+ and F_+ are semisimple. ■

Remark 11.10. We note that the partial converse given in Theorem 11.9(b) is the best we can do. If one of the cones is trivial, then the outcome depends on the other cone. Indeed, let $E, F \neq \{0\}$ with convex cones $E_+ \subseteq E$, $F_+ \subseteq F$, such that $E_+ = \{0\}$ and F_+ is not semisimple. Then $E_+ \otimes^\pi F_+ = \{0\}$, which is semisimple, but $E_+ \otimes^\varepsilon F_+$ is not semisimple by Theorem 10.22.

More can be said if we choose the cone beforehand. The injective cone is already weakly closed with respect to any reasonable dual, so Theorem 10.22 tells us exactly when $E_+ \otimes^\varepsilon F_+$ is semisimple. For the projective cone, we obtain necessary and sufficient criteria very similar to those in Theorem 9.10.

Corollary 11.11. *Let $\langle E, E' \rangle$, $\langle F, F' \rangle$ be dual pairs, let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones, and let G be a reasonable dual of $E \otimes F$. Then $E_+ \otimes^\pi F_+$ is G-semisimple if and only if $E_+ = \{0\}$, or $F_+ = \{0\}$, or both E_+ and F_+ are semisimple.*

Proof. If $E_+ = \{0\}$ or $F_+ = \{0\}$, then $E_+ \otimes^\pi F_+ = \{0\}$, which is semisimple. The rest follows from Theorem 11.9. ■

Remark 11.12. Barring corner cases, we find that $E_+ \otimes^\pi F_+$ is semisimple if and only if $E_+ \otimes^\varepsilon F_+$ is a proper cone. It is tempting to conjecture that the projective cone is always dense in the injective cone. For locally convex lattices, Birnbaum [Bir76, Proposition 3] found a positive answer, but in general this is far from being true. Counterexamples have been known for a long time (e.g. [Bir76, Example following Proposition 3]; [BL75, Proposition 3.1]). Very recently, Aubrun, Lami, Palazuelos and Plávala [ALPP21] proved that this fails for all closed, proper, generating cones in finite-dimensional spaces, unless at least one of the cones is a simplex cone. We will prove a large class of special cases of this result in Chapter 13.

Remark 11.13. Fremlin [Fre72] developed a theory of tensor products of Archimedean Riesz spaces, which was further developed by Grobler and Labuschagne [GL88], and van Gaans and Kalauch [GK10] to a theory of tensor products of Archimedean cones. In this setting, the challenge is to extend the projective cone to a proper Archimedean cone. In [GK10], van Gaans and Kalauch showed that the projective tensor product of

two generating Archimedean cones is always contained in a proper Archimedean cone (see [GK10, Lemma 4.2]).

Our results are parallel to this. If the given cones $E_+ \subseteq E$ and $F_+ \subseteq F$ are not only Archimedean but also closed in some locally convex topology (this is a stronger assumption), then their projective tensor product is contained in a closed (hence Archimedean) proper cone. In other words, we start with a stronger assumption, and end up with a stronger conclusion.

The preceding results are no substitute for the methods developed in [GK10]. For example, the space $L^p[0,1]$ with $p \in (0,1)$ does not admit a non-trivial positive linear functional, so here we have an Archimedean cone which fails to be semisimple in a rather dramatic way. Consequently, our results fail to prove that the projective tensor product $L^p_+[0,1] \otimes^\pi L^p_+[0,1]$ is contained in a proper Archimedean cone, which we know to be true by the results of [GK10]. (In fact, since $L^p_+[0,1]$ is a lattice cone, this follows already from Fremlin's original result [Fre72, Theorem 4.2]).

11.4 Semisimplicity of reasonable crosscones in completed locally convex tensor products

In the completed setting, semisimplicity turns out to be more subtle. This is because there is one additional requirement for the injective cone to be proper: not only do $\overline{E_+}^w$ and $\overline{F_+}^w$ need to be proper, but the natural map $E \tilde{\otimes}_\alpha F \to E \tilde{\otimes}_\varepsilon F$ must be injective. (See Corollary 10.23.) This leads to the following analogue of Theorem 11.9.

Theorem 11.14. *Let E, F be complete locally convex spaces, $E_+ \subseteq E$, $F_+ \subseteq F$ convex cones, α a compatible locally convex topology on $E \otimes F$, and $\mathcal{K} \subseteq E \tilde{\otimes}_\alpha F$ a reasonable crosscone.*

(a) If E_+ and F_+ are semisimple and if $E \tilde{\otimes}_\alpha F \to E \tilde{\otimes}_\varepsilon F$ is injective, then $\mathcal{K}$ is semisimple.

(b) If $E_+ \neq \{0\}$ and $F_+ \neq \{0\}$, and if $\mathcal{K}$ is semisimple, then E_+ and F_+ are semisimple.

Proof.

(a) It follows from the assumptions and Corollary 10.23 that the injective cone $E_+ \tilde{\otimes}^\varepsilon_\alpha F_+$ is proper, so $\mathcal{K}$ is contained in a closed proper cone.

(b) If $\mathcal{K}$ is semisimple, then in particular $\mathcal{K} \cap (E \otimes F)$ is semisimple, so the result follows from Theorem 11.9(b). ■

The gap between the necessary and sufficient conditions in Theorem 11.14 is even larger than it was in Theorem 11.9. We show that this gap is related to the approximation property. For simplicity, we restrict our attention to Banach spaces.

We recall some generalities. Let α be a finitely generated tensor norm, then we say (following [DF93, §21.7]) that a Banach space E has the *α-approximation property* if for all Banach spaces F the natural map $E \tilde{\otimes}_\alpha F \to E \tilde{\otimes}_\varepsilon F$ is injective. The π-approximation property (where π denotes the projective tensor norm) is simply called

the *approximation property*. If a Banach space E has the approximation property, then E also has the α-approximation property for every finitely generated tensor norm α (see [DF93, Proposition 17.20]).

Some tensor norms α have the property that every Banach space has the α-approximation property (and therefore $E \tilde{\otimes}_\alpha F \to E \tilde{\otimes}_\varepsilon F$ is always injective). One of these is the injective tensor norm ε, for obvious reasons. More generally, this is true for every totally accessible tensor norm α; see [DF93, Proposition 21.7(2)]. This includes all tensor norms which are (left and right) injective; see [DF93, Proposition 21.1(3)].

Corollary 11.15. *Let E and F be Banach spaces, let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones, and let α be a finitely generated tensor norm. If E or F has the α-approximation property, then the projective cone $E_+ \tilde{\otimes}^\pi_\alpha F_+ \subseteq E \tilde{\otimes}_\alpha F$ is semisimple if and only if $E_+ = \{0\}$, or $F_+ = \{0\}$, or both E_+ and F_+ are semisimple.*

Proof. The α-approximation property guarantees that $E \tilde{\otimes}_\alpha F \to E \tilde{\otimes}_\varepsilon F$ is injective. If $E_+ = \{0\}$ or $F_+ = \{0\}$, then $E_+ \tilde{\otimes}^\pi_\pi F_+ = \{0\}$. The other cases follow from Theorem 11.14. ■

The proofs of Corollary 11.11 and Corollary 11.15 rely on the injective cone to draw conclusions about the projective cone. However, in general these two can be far apart (see Remark 11.12). If the map $E \tilde{\otimes}_\alpha F \to E \tilde{\otimes}_\varepsilon F$ is not injective, then the injective cone $E_+ \tilde{\otimes}^\varepsilon_\alpha F_+$ is not proper, but that does not mean that the projective cone $E_+ \tilde{\otimes}^\pi_\alpha F_+$ cannot be semisimple. This leaves open the following interesting question, to which we do not know the answer:

Question 11.16. *Let E, F be real Banach spaces, and let $E_+ \subseteq E$, $F_+ \subseteq F$ be closed proper cones. Is the projective cone $E_+ \tilde{\otimes}^\pi_\pi F_+$ in the completed projective tensor product $E \tilde{\otimes}_\pi F$ necessarily contained in a closed proper cone?*

Equivalently: if the positive continuous linear functionals separate points on E and F, then do the positive continuous bilinear forms $E \times F \to \mathbb{R}$ separate points on $E \tilde{\otimes}_\pi F$?

By Corollary 11.11, the positive continuous bilinear forms separate points on $E \otimes_\pi F$, but that is not enough. Furthermore, if E or F has the approximation property, then the positive bilinear forms of rank one already separate points on $E \tilde{\otimes}_\pi F$, but this technique does not work in the absence of the approximation property.

Chapter 12

Basic additional properties in the finite-dimensional case

In this chapter, we give an overview of the most important additional properties in the finite-dimensional case. This will be used extensively in the next chapter, where we give many examples where the projective cone is not dense in the injective cone.

This chapter is based on Chapter 6 of [Dob20b].

Introduction

In the previous chapters, we studied tensor products of convex cones in general (possibly infinite-dimensional) real vector spaces. In that setting, not many results had been known in the literature, and several basic questions had been unanswered. In the finite-dimensional setting, the situation is very different. In a different part of the literature, completely separate from the functional analysis literature, questions around tensor products of closed, proper and generating cones have been studied by many authors in a variety of different fields, such as linear algebra, operator theory, geometry, approximation theory, and theoretical physics. (For a comprehensive overview of these connections, see §7.1.)

In this chapter, we give a brief overview of the most important additional properties in the finite-dimensional case. We give new, streamlined proofs of several known results, and we extend them to general convex cones in finite-dimensional spaces (i.e. cones which are not necessarily closed, proper, or generating). In §12.2, we show that the projective and injective cones can be interpreted as certain cones of positive operators, at least when E_+ and F_+ are closed. In §12.3, we show that the closure $\overline{E_+ \otimes^\pi F_+}$ of the projective cone $E_+ \otimes^\pi F_+$ is equal to the projective cone $\overline{E_+} \otimes^\pi \overline{F_+}$. Finally, in §12.4, we look more closely at the concept of retracts, which was already covered briefly in §8.3. Here we prove some basic properties of retracts, and we provide many examples of retracts in standard cones, which we will use in the next chapter.

12.1 Additional notation

We follow notation from Chapter 8; see also the glossary of notation on page 201.

In this chapter (and the next), all vector spaces will be finite-dimensional. Recall that the *lineality space* of a convex cone $\mathcal{K} \subseteq E$ is the subspace $\lin(\mathcal{K}) = \mathcal{K} \cap -\mathcal{K}$. We say that $\mathcal{K}$ is *proper* if $\lin(\mathcal{K}) = \{0\}$, and *generating* if $\mathcal{K} - \mathcal{K} = E$.

The (*algebraic*) *dual cone* $\mathcal{K}^* \subseteq E^*$ is the set of positive linear functionals:

$$\mathcal{K}^* := \big\{\varphi \in E^* \,:\, \langle x, \varphi\rangle \geq 0 \text{ for all } x \in \mathcal{K}\big\}.$$

This is a closed cone in the (finite-dimensional) space E^*, and the natural isomorphism $E^{**} \cong E$ identifies the double dual cone $\mathcal{K}^{**}$ with the closure $\overline{\mathcal{K}}$.

If $\mathcal{K}$ is a convex cone, then a *base of* $\mathcal{K}$ is a convex subset $\mathcal{B} \subseteq \mathcal{K} \setminus \{0\}$ such that each $x \in \mathcal{K} \setminus \{0\}$ can be written uniquely as $x = \lambda b$ with $\lambda > 0$ and $b \in \mathcal{B}$. If $\mathcal{K}$ is generating, then the bases of $\mathcal{K}$ are in bijective correspondence with the strictly $\mathcal{K}$-positive linear functionals on E (see [AT07, Theorem 1.47]). Furthermore, every closed proper cone in a finite-dimensional space admits a compact base (e.g. [AT07, Corollary 3.8]).

We say that a convex cone $E_+ \subseteq E$ is a *simplex cone* (or *Yudin cone*) if it is generated by a basis of E, or equivalently, if every base of E_+ is a simplex. A simplex cone turns E into a Dedekind complete Riesz space (see [AT07, Theorem 3.17]). Furthermore, a cone in a finite-dimensional space is a simplex cone if and only if it is a closed lattice cone (see [AT07, Theorem 3.21]).

We fix notation for a number of standard cones. For $n \geq 1$, we let $\mathcal{L}^n \subseteq \mathbb{R}^n$ denote the n-dimensional *second-order cone* (or *Lorentz cone*, or *ice cream cone*),

$$\mathcal{L}^n := \big\{(x_1, \ldots, x_n) \in \mathbb{R}^n \,:\, \sqrt{x_1^2 + \cdots + x_{n-1}^2} \leq x_n\big\}.$$

(By convention, $\mathcal{L}^1$ is just the standard cone $\mathbb{R}_+ \subseteq \mathbb{R}$.) Furthermore, let $\mathcal{S}^n \subseteq \mathbb{R}^{n\times n}$ and $\mathcal{H}^n \subseteq \mathbb{C}^{n\times n}$ denote the spaces of real symmetric and complex hermitian $n \times n$ matrices, respectively. We denote the respective positive semidefinite cones by $\mathcal{S}^n_+$ and $\mathcal{H}^n_+$:

$$\mathcal{S}^n_+ := \{n \times n \text{ real positive semidefinite matrices}\};$$
$$\mathcal{H}^n_+ := \{n \times n \text{ complex positive semidefinite matrices}\}.$$

Recall that $\mathcal{S}^2_+$ is isomorphic with the Lorentz cone $\mathcal{L}^3$, for instance via the isomorphism

$$\mathcal{S}^2 \to \mathbb{R}^3, \qquad \begin{pmatrix} a & b \\ b & c \end{pmatrix} \mapsto (a - c, 2b, a + c).$$

(Use that $A \in \mathcal{S}^2$ is positive semidefinite if and only if $\tr(A) \geq 0$ and $\det(A) \geq 0$.)

12.2 Simplex-factorable positive linear maps

For the remainder of this chapter (and the next), it will be convenient to reformulate questions regarding the projective and injective cones in terms of positive linear operators.

If E and F are preordered by convex cones $E_+ \subseteq E$, $F_+ \subseteq F$, then a positive linear operator $T : E \to F$ is called *separable*[1] if it can be written as $T = \sum_{i=1}^{k} \varphi_i \otimes y_i$, where $\varphi_1, \dots, \varphi_k \in E_+^*$ are positive linear functionals and $y_1, \dots, y_k \in F_+$ are positive elements.

If G is another finite-dimensional real vector space, preordered by a convex cone $G_+ \subseteq G$, then we say that a positive linear map $T : E \to F$ *factors through* G is there exist positive linear maps $R : E \to G$ and $S : G \to F$ such that $T = S \circ R$. We say that T *factors through a simplex cone*[2] if it factors through some $\mathbb{R}^n$, ordered by the standard cone $\mathbb{R}^n_{\geq 0}$.

Proposition 12.1. *A positive linear map $T : E \to F$ is separable if and only if it factors through a simplex cone $\mathbb{R}^n_{\geq 0}$. If this is the case, then one may take $n \leq \dim(E) \times \dim(F)$.*

Proof. "$\Longrightarrow$". Write $T = \sum_{i=1}^{k} \varphi_i \otimes y_i$ with $\varphi_1, \dots, \varphi_k \in E_+^*$, $y_1, \dots, y_k \in F_+$. By Carathéodory's theorem for cones (see e.g. [Roc70, Corollary 17.1.2]), we may assume without loss of generality that $k \leq \dim(E^* \otimes F) = \dim(E) \times \dim(F)$. Now define $R : E \to \mathbb{R}^k$ and $S : \mathbb{R}^k \to F$ by

$$R(x) := (\varphi_1(x), \dots, \varphi_k(x));$$
$$S(\lambda_1, \dots, \lambda_k) := \lambda_1 y_1 + \cdots + \lambda_k y_k.$$

Then R and S are positive ($\mathbb{R}^k$ equipped with the standard cone $\mathbb{R}^k_{\geq 0}$), and $T = S \circ R$.

"$\Longleftarrow$". Suppose that T factors as

$$E \xrightarrow{\ R\ } \mathbb{R}^n \xrightarrow{\ S\ } F,$$

with R and S positive ($\mathbb{R}^n$ equipped with the standard cone $\mathbb{R}^n_{\geq 0}$). Let $e_1, \dots, e_n \in \mathbb{R}^n$ denote the standard basis of $\mathbb{R}^n$, and $e_1^*, \dots, e_n^*$ the corresponding dual basis. Define $\varphi_1, \dots, \varphi_n \in E^*$ and $y_1, \dots, y_n \in F$ by setting $\varphi_i := e_i^* \circ R$ and $y_i := S(e_i)$. Then $\varphi_1, \dots, \varphi_n \in E_+^*$, $y_1, \dots, y_n \in F_+$, and $T = \sum_{i=1}^{n} \varphi_i \otimes y_i$. ■

Corollary 12.2. *If E_+ and F_+ are closed, then:*

(a) $E_+^* \otimes^{\varepsilon} F_+$ is the set of all positive linear maps $E \to F$;

(b) $E_+^* \otimes^{\pi} F_+$ is the set of all positive linear maps $E \to F$ that factor through a simplex cone.

Proof. It is well known that $E_+^* \otimes^{\varepsilon} F_+$ can be identified with the cone of linear maps $T : E \to F$ satisfying $T[\overline{E_+}] \subseteq \overline{F_+}$ (see Remark 10.2). Since E_+ and F_+ are closed, these are simply the positive linear maps $E \to F$.

Under this identification, it is clear from the definition that $E_+^* \otimes^{\pi} F_+$ is the subset of separable positive linear maps $E \to F$. By Proposition 12.1 these are precisely the positive linear maps $E \to F$ that factor through a simplex cone. ■

[1] The terminology was introduced by other authors in connection with quantum theory (e.g. [Hil08, ALP19]).

[2] More accurately, we should say that *T factors through a finite-dimensional Archimedean Riesz space*, but this is too wordy.

12.3 The closure of the projective cone

In this section, we prove that the closure of the projective cone $E_+ \otimes^\pi F_+$ is equal to the projective tensor product of $\overline{E_+}$ and $\overline{F_+}$. In particular, the projective tensor product of closed convex cones is closed. In the case where E_+ and F_+ are closed, proper, and generating, this was already established by Tam [Tam77b].

The proof is carried out in three steps. First we prove the result for closed proper cones, thereby giving another proof of the aforementioned result by Tam. Secondly, we extend this to all closed convex cones by decomposing a closed convex cone as the sum of a closed proper cone and a subspace. After this, it will be relatively simple to deduce the general formula for the closure of the projective cone.

The projective tensor product of closed proper cones

Recall that a *base* (of E_+) is a convex subset $\mathcal{B} \subseteq E_+$ with $0 \notin \mathcal{B}$ such that every $x \in E_+ \setminus \{0\}$ can be written uniquely as $x = \lambda b$ with $\lambda > 0$ and $b \in \mathcal{B}$. A key property is that a subset $\mathcal{B} \subseteq E_+$ is a base if and only if there is a strictly positive linear functional $f : E \to \mathbb{R}$ such that $\mathcal{B} = \{x \in E_+ : f(x) = 1\}$; see e.g. [AT07, Theorem 1.47].

Not every proper cone has a base. However, every closed proper cone in a finite-dimensional space has a compact base (e.g. [AT07, Corollary 3.8]), and conversely the convex cone generated by a compact convex set $S \subseteq \mathbb{R}^n \setminus \{0\}$ is a closed proper cone (e.g. [AT07, Lemma 3.12]).

Proposition 12.3. *Let E, F be real vector spaces, and let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones having bases $\mathcal{B}_{E_+} \subseteq E_+$, $\mathcal{B}_{F_+} \subseteq F_+$. Then* $\operatorname{conv}(\mathcal{B}_{E_+} \otimes \mathcal{B}_{F_+})$ *is a base of $E_+ \otimes^\pi F_+$.*

Proof. Let $f : E \to \mathbb{R}$ and $g : F \to \mathbb{R}$ be strictly positive linear functionals such that $\mathcal{B}_{E_+} = \{x \in E_+ : f(x) = 1\}$ and $\mathcal{B}_{F_+} = \{y \in F_+ : g(y) = 1\}$. Then $f \otimes g$ is a strictly positive linear functional on $E \otimes F$ (with respect to the projective cone), and we have

$$\mathcal{B}_{E_+} \otimes \mathcal{B}_{F_+} \subseteq \{z \in E_+ \otimes^\pi F_+ : (f \otimes g)(z) = 1\}.$$

Since the right-hand side is convex, it follows that

$$\operatorname{conv}(\mathcal{B}_{E_+} \otimes \mathcal{B}_{F_+}) \subseteq \{z \in E_+ \otimes^\pi F_+ : (f \otimes g)(z) = 1\}. \tag{12.4}$$

On the other hand, every non-zero element of $E_+ \otimes^\pi F_+$ can be written as a positive multiple of an element in $\operatorname{conv}(\mathcal{B}_{E_+} \otimes \mathcal{B}_{F_+})$, so we must have equality in (12.4). ■

Corollary 12.5 ([Tam77b])**.** *If $E_+ \subseteq E$, $F_+ \subseteq F$ are closed proper cones, then $E_+ \otimes^\pi F_+$ is also a closed proper cone.*

Proof. Choose compact bases $\mathcal{B}_{E_+} \subseteq E_+$ and $\mathcal{B}_{F_+} \subseteq F_+$. Then, by Proposition 12.3, $\operatorname{conv}(\mathcal{B}_{E_+} \otimes \mathcal{B}_{F_+})$ is a base for $E_+ \otimes^\pi F_+$. In particular, $0 \notin \operatorname{conv}(\mathcal{B}_{E_+} \otimes \mathcal{B}_{F_+})$.

The natural map $E \times F \to E \otimes F$ is continuous, so $\mathcal{B}_{E_+} \otimes \mathcal{B}_{F_+}$ is compact in $E \otimes F$. Since the convex hull of a compact set in $\mathbb{R}^n$ is also compact (e.g. [Rud91, Theorem

3.20(d)]), it follows that $\mathrm{conv}(\mathcal{B}_{E_+} \otimes \mathcal{B}_{F_+})$ is compact. Thus, $E_+ \otimes^\pi F_+$ is generated by a compact convex set not containing 0, so it follows that $E_+ \otimes^\pi F_+$ is a closed proper cone. ■

Remark 12.6. The proof of Corollary 12.5 shows directly that $E_+ \otimes^\pi F_+$ is proper whenever E_+ and F_+ are closed proper cones. As such, this provides yet another way to prove that the projective tensor product of proper cones is always proper, in addition to the different ways discussed in Remark 9.12.

The projective tensor product of closed convex cones

In order to extend Corollary 12.5 to all closed convex cones, we decompose each of E_+ and F_+ as the sum of a closed proper cone and a subspace. The (straightforward) proof of the following classical result is omitted.

Proposition 12.7. *Let E be a finite-dimensional, and let $E_+ \subseteq E$ be a closed convex cone. Let $\mathrm{lin}(E_+) := E_+ \cap -E_+$ be the lineality space of E_+, and let $\mathrm{lin}(E_+)^\perp$ be any complementary subspace of $\mathrm{lin}(E_+)$. Then $\mathrm{lin}(E_+)^\perp_+ := \mathrm{lin}(E_+)^\perp \cap E_+$ is a closed proper cone, and one has*

$$E_+ = \mathrm{lin}(E_+) + \mathrm{lin}(E_+)^\perp_+.$$

Conversely, the sum of a closed proper cone and a subspace need not be closed. (Example: let E_+ be the cone generated by $\{(x, y, 1) \in \mathbb{R}^3 : (x-1)^2 + y^2 \leq 1\}$, and $X := \mathrm{span}\{(0, 0, 1)\} \subseteq \mathbb{R}^3$.) However, we have the following partial converse of Proposition 12.7.

Proposition 12.8. *Let E be a finite-dimensional, let $E_+ \subseteq E$ be a closed convex cone, and let $X \subseteq E$ be a subspace. If $\mathrm{span}(E_+) \cap X = \{0\}$, then $E_+ + X$ is a closed convex cone.*

Proof. Extend $\mathrm{span}(E_+)$ to a complementary subspace $X^\perp$ of X. Let $P : E \to X^\perp$ be the projection $x + x^\perp \mapsto x^\perp$. Then P is continuous, and $E_+ + X = P^{-1}[E_+]$, so $E_+ + X$ is closed. ■

The preceding propositions give us a way to decompose the cones and later put them back together. To see what happens when we lift the pieces separately, we use the following observation.

Proposition 12.9. *Let E, F be real vector spaces, and let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones. If at least one of E_+ and F_+ is a subspace, then $E_+ \otimes^\pi F_+$ is a subspace as well.*

Proof. A convex cone G_+ is a subspace precisely when one has $s \in G_+$ if and only if $-s \in G_+$. This property is preserved by the projective tensor product. ■

We can now extend Corollary 12.5 to all closed convex cones.

Theorem 12.10. *Let E, F be finite-dimensional, and let $E_+ \subseteq E$, $F_+ \subseteq F$ be closed convex cones. Then $E_+ \otimes^\pi F_+$ is closed as well.*

Proof. Choose complementary subspaces $\operatorname{lin}(E_+)^\perp \subseteq E$ and $\operatorname{lin}(F_+)^\perp \subseteq F$ of $\operatorname{lin}(E_+)$ and $\operatorname{lin}(F_+)$. Then, by Proposition 12.7, we have $E_+ = \operatorname{lin}(E_+) + \operatorname{lin}(E_+)^\perp_+$ and $F_+ = \operatorname{lin}(F_+) + \operatorname{lin}(F_+)^\perp_+$, with $\operatorname{lin}(E_+)^\perp_+$ and $\operatorname{lin}(F_+)^\perp_+$ closed proper cones. It follows that

$$E_+ \otimes^\pi F_+ = \overbrace{\big(\operatorname{lin}(E_+) \otimes^\pi \operatorname{lin}(F_+)\big)}^{\text{subspace}} + \overbrace{\big(\operatorname{lin}(E_+) \otimes^\pi \operatorname{lin}(F_+)^\perp_+\big)}^{\text{subspace}} + \overbrace{\big(\operatorname{lin}(E_+)^\perp_+ \otimes^\pi \operatorname{lin}(F_+)\big)}^{\text{subspace}} + \underbrace{\big(\operatorname{lin}(E_+)^\perp_+ \otimes^\pi \operatorname{lin}(F_+)^\perp_+\big)}_{\text{closed proper cone}}.$$

(The first three terms are subspaces by Proposition 12.9; the fourth is a closed proper cone by Corollary 12.5.) The three subspaces in the preceding formula are contained in the subspace $(\operatorname{lin}(E_+)\otimes\operatorname{lin}(F_+))+(\operatorname{lin}(E_+)\otimes\operatorname{lin}(F_+)^\perp)+(\operatorname{lin}(E_+)^\perp\otimes\operatorname{lin}(F_+))$, whereas $\operatorname{lin}(E_+)^\perp_+ \otimes^\pi \operatorname{lin}(F_+)^\perp_+$ is a closed proper cone contained within $\operatorname{lin}(E_+)^\perp \otimes \operatorname{lin}(F_+)^\perp$. These containing subspaces are complementary, so it follows from Proposition 12.8 that $E_+ \otimes^\pi F_+$ is closed. ∎

Remark 12.11. We should point out that nothing like Theorem 12.10 is true in the infinite-dimensional setting. In fact, the projective tensor product of closed proper cones in Banach spaces might not even be Archimedean (see e.g. [PTT11, Remark 3.12]).

The closure of the projective cone; duality

Using Theorem 12.10, it is now relatively easy to prove the following.

Theorem 12.12. *Let E and F be finite-dimensional real vector spaces, and let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones. Then the closure of the projective cone $E_+ \otimes^\pi F_+$ is the projective cone $\overline{E_+} \otimes^\pi \overline{F_+}$.*

Proof. "$\supseteq$". Given $x \in \overline{E_+}$, $y \in \overline{F_+}$, choose sequences $\{x_n\}_{n=1}^\infty$ and $\{y_n\}_{n=1}^\infty$ in E_+ and F_+ converging to x and y, respectively. Then $x \otimes y = \lim_{n\to\infty} x_n \otimes y_n \in \overline{E_+ \otimes^\pi F_+}$. (Alternatively, use Proposition 11.8.)

"$\subseteq$". Evidently, $E_+ \otimes^\pi F_+ \subseteq \overline{E_+} \otimes^\pi \overline{F_+}$. By Theorem 12.10, $\overline{E_+} \otimes^\pi \overline{F_+}$ is closed, so we also have $\overline{E_+ \otimes^\pi F_+} \subseteq \overline{E_+} \otimes^\pi \overline{F_+}$. ∎

Other consequences of Theorem 12.10 include the following.

Corollary 12.13. *Let E, F be finite-dimensional, and let $E_+ \subseteq E$, $F_+ \subseteq F$ be convex cones. Then:*

(a) $E_+ \otimes^\pi F_+$ is dense in $E_+ \otimes^\varepsilon F_+$ if and only if $\overline{E_+} \otimes^\pi \overline{F_+} = \overline{E_+} \otimes^\varepsilon \overline{F_+}$;

(b) $(E_+ \otimes^\varepsilon F_+)^* = E_+^* \otimes^\pi F_+^*$.

Proof.

(a) It is clear from the definition that $E_+ \otimes^\varepsilon F_+ = \overline{E_+} \otimes^\varepsilon \overline{F_+}$, and that this cone is always closed. Thus, the conclusion follows immediately from Theorem 12.12.

(b) We have $(E_+^* \otimes^\pi F_+^*)^* = E_+^{**} \otimes^\varepsilon F_+^{**} = \overline{E_+} \otimes^\varepsilon \overline{F_+}$. Using again that $E_+ \otimes^\varepsilon F_+ = \overline{E_+} \otimes^\varepsilon \overline{F_+}$, we find that $(E_+^* \otimes^\pi F_+^*)^* = E_+ \otimes^\varepsilon F_+$. By Theorem 12.10, $E_+^* \otimes^\pi F_+^*$ is closed, so the result follows by duality. ■

In other words, if the spaces are finite-dimensional and the cones are closed, then we have full duality between the projective and injective cones.

12.4 Retracts

In Chapter 13, we will look at the problem of determining whether or not $E_+ \otimes^\pi F_+$ and $E_+ \otimes^\varepsilon F_+$ coincide. This problem can sometimes be reduced to lower dimensional spaces by using retracts.

Let (F, F_+) be a finite-dimensional preordered vector space. Then a subspace $E \subseteq F$ is called an *order retract* if there exists a positive projection $F \to E$. More generally, another preordered space (G, G_+) is *isomorphically an order retract* if there exist positive linear maps $T : G \to F$ and $S : F \to G$ such that $S \circ T = \mathrm{id}_G$. Note that in this case T is automatically bipositive (i.e. a pullback) and S is automatically a pushforward, and $\mathrm{ran}(T) \subseteq F$ is a retract of F which is order isomorphic to G.

For simplicity, we shall omit the word *order* when talking about retracts, for there is minimal chance of confusion with other types of retracts (e.g. from topology).

Although retracts do not appear to be a very common notion in the theory of ordered vector spaces, some of the results from this section were discovered independently by Aubrun, Lami and Palazuelos [ALP19].

Remark 12.14. Some basic properties of retracts:

(a) if E is a retract of (F, F_+) and F_+ is a proper cone, then $E_+ := E \cap F_+$ is a proper cone (after all, E_+ is a subcone);

(b) if E is a retract of (F, F_+) and F_+ is generating, then E_+ is generating in E (after all, there exists a surjective positive operator $F \to E$);

(c) if (E, E_+) is isomorphically a retract of (F, F_+), and if (F, F_+) is isomorphically a retract of (G, G_+), then (E, E_+) is isomorphically a retract of (G, G_+).

(d) If (E, E_+) is isomorphically a retract of $(F.F_+)$, then (E^*, E_+^*) is isomorphically a retract of (F^*, F_+^*). After all, if $T : E \to F$ and $S : F \to E$ are positive linear maps with $\mathrm{id}_E = S \circ T$, then $S^* : E^* \to F^*$ and $T^* : F^* \to E^*$ are positive linear maps with

$$\mathrm{id}_{E^*} = (\mathrm{id}_E)^* = (S \circ T)^* = T^* \circ S^*.$$

By Remark 12.14(b), if $F_+ \subseteq F$ is generating, then a retract $E \subseteq F$ is uniquely determined by its positive part $E_+ := E \cap F_+$, so instead of saying that E *is a retract of* (F, F_+) we will simply say that E_+ *is a retract of* F_+.

Example 12.15. We present some examples of retracts.

(a) If E is finite-dimensional and E_+ is a closed and proper convex cone, then every (not necessarily extremal) ray in E_+ is a retract. Indeed, let $x_0 \in E_+ \setminus \{0\}$ be arbitrary, and let $\varphi_0 \in E_+^*$ be a strictly positive linear functional. Then $\frac{1}{\varphi_0(x_0)} \cdot \varphi_0 \otimes x_0$ defines a positive projection onto $\operatorname{span}(x_0)$.

(b) If $n \leq m$, then $\mathbb{R}^n_{\geq 0}$ is a retract of $\mathbb{R}^m_{\geq 0}$, for instance via the maps $\mathbb{R}^n \to \mathbb{R}^m$ and $\mathbb{R}^m \to \mathbb{R}^n$ given respectively by padding with zeroes and projecting onto the first n coordinates. Although (a) shows that these are not the only retracts, we will show in Lemma 13.5 that every retract of a simplex cone is once again a simplex cone.

(c) In the same manner, if $n \leq m$, then $\mathcal{S}^n_+$ is a retract of $\mathcal{S}^m_+$, and $\mathcal{H}^n_+$ is a retract of $\mathcal{H}^m_+$.

(d) If $n \leq m$, then $\mathcal{L}^n$ is a retract of $\mathcal{L}^m$ via the maps $T : \mathbb{R}^n \to \mathbb{R}^m$, $S : \mathbb{R}^m \to \mathbb{R}^n$ given by

$$T(x_1, \ldots, x_n) = (x_1, \ldots, x_{n-1}, 0, \ldots, 0, x_n);$$
$$S(y_1, \ldots, y_m) = (y_1, \ldots, y_{n-1}, y_m).$$

(e) $\mathbb{R}^n_{\geq 0}$ is a retract of $\mathcal{S}^n_+$ via the map $T : \mathbb{R}^n \to \mathcal{S}^n$ that maps x to the diagonal matrix whose entries are specified by x, and the map $S : \mathcal{S}^n \to \mathbb{R}^n$ that maps A to the diagonal of A.

(f) $\mathcal{S}^n_+$ is a retract of $\mathcal{H}^n_+$, via the maps $T : \mathcal{S}^n \to \mathcal{H}^n$, $A \mapsto A$ and $S : \mathcal{H}^n \to \mathcal{S}^n$, $A \mapsto \frac{1}{2}(A + \overline{A})$.

(g) $\mathcal{H}^n_+$ is a retract of $\mathcal{S}^{2n}_+$, via the maps $T : \mathcal{H}^n \to \mathcal{S}^{2n}$ and $S : \mathcal{S}^{2n} \to \mathcal{H}^n$ given by

$$T(A + iB) = \begin{pmatrix} A & -B \\ B & A \end{pmatrix}, \qquad S \begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix} = \frac{1}{2}(A_1 + A_4) + \frac{i}{2}(A_3 - A_2). \quad \triangle$$

A more advanced example occurs in polyhedral cones. If E_+ is a proper and generating polyhedral cone with extremal directions $\{x_0, \ldots, x_k\}$, then a *vertex figure at* x_0 is a subcone of the form $E_+ \cap \ker(\varphi_0)$, where $\varphi_0 \in E^*$ is a linear form such that $\varphi_0(x_0) < 0$ and $\varphi_0(x_i) > 0$ for all $i > 0$. Vertex figures are combinatorially dual to facets (e.g. [Brø83, Theorem 11.5]).

Proposition 12.16. *Let E be finite-dimensional and let $E_+ \subseteq E$ be a proper and generating polyhedral cone. Then every vertex figure and every facet of E_+ is a retract.*

Proof. Let $x_0, \ldots, x_k$ be the extremal directions of E_+, and suppose that $\varphi_0 \in E^*$ defines a vertex figure at x_0 (i.e. $\varphi_0(x_0) < 0$ and $\varphi_0(x_i) > 0$ for all $i > 0$). We show that $E_+ \cap \ker(\varphi_0)$ is a retract.

By scaling, we may assume without loss of generality that $\varphi_0(x_0) = -1$. Define $P_{\varphi_0} : E \to E$ by $y \mapsto y + \varphi_0(y)x_0$. We show that P_{φ_0} is a positive projection onto $\ker(\varphi_0)$. For all $y \in E$ we have

$$\varphi_0(P_{\varphi_0}(y)) = \varphi_0(y) + \varphi_0(y)\varphi_0(x_0) = \varphi_0(y) - \varphi_0(y) = 0,$$

which shows that $\operatorname{ran}(P_{\varphi_0}) \subseteq \ker(\varphi_0)$. Furthermore, if $y \in \ker(\varphi_0)$, then $P_{\varphi_0}(y) = y + 0 = y$, so P_{φ_0} is a projection onto $\ker(\varphi_0)$. To prove positivity, it suffices to show that $P_{\varphi_0}(x_i) \in E_+$ for all i. We distinguish two cases:

- For $i = 0$, we have $P_{\varphi_0}(x_0) = x_0 + \varphi_0(x_0)x_0 = x_0 - x_0 = 0 \in E_+$.
- For $i > 0$, we have $\varphi_0(x_i) > 0$, hence $P_{\varphi_0}(x_i) = x_i + \varphi_0(x_i)x_0 \in E_+$.

This shows that every vertex figure is a retract. Additionally, note that $\ker(P_{\varphi_0}) = \operatorname{span}(x_0)$; this will be used in the second part of the proof.

Now let $M \subseteq E_+$ be a facet. Then M corresponds with an extremal direction $\psi_0 \in E_+^* \setminus \{0\}$ of the dual cone, in such a way that $M^\perp = \operatorname{span}(\psi_0)$. Choose a vertex figure $N \subseteq E_+^*$ at ψ_0. The preceding argument shows that there are positive linear maps $T : \operatorname{span}(N) \hookrightarrow E^*$ and $S : E^* \twoheadrightarrow \operatorname{span}(N)$ such that $\mathrm{id}_{\operatorname{span}(N)} = ST$. Furthermore, the construction gives us the additional property that $\ker(S) = \operatorname{span}(\psi_0)$. Dualizing the retract (see Remark 12.14(d)) shows that $\operatorname{span}(N)^*$ is isomorphically a retract of E, by means of the maps $S^* : \operatorname{span}(N)^* \hookrightarrow E$ and $T^* : E \twoheadrightarrow \operatorname{span}(N)^*$. Since $\operatorname{ran}(S^*) = {}^\perp\ker(S) = {}^\perp\{\psi_0\} = \operatorname{span}(M)$, this shows that M is a retract of E_+. ■

Remark 12.17. In fact, retractions give a *geometric* duality between facets and vertex figures (in addition to the well-known *combinatorial* duality). If $M \subseteq E_+$ is a facet corresponding to the extremal direction $\psi_0 \in E_+^* \setminus \{0\}$ of the dual cone, then one can show that:

- every vertex figure at ψ_0 admits a *unique* positive projection (namely, the one from the proof of Proposition 12.16);
- there is a bijective correspondence between vertex figures at ψ_0 and positive projections $E \to \operatorname{span}(M)$ that map every element of $E_+ \setminus M$ in the relative interior of M.

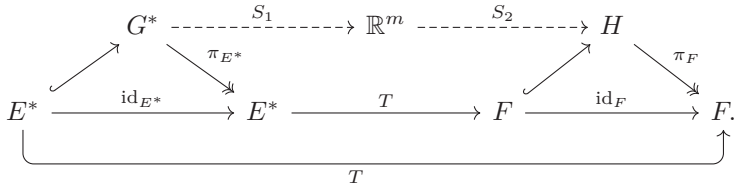
As we have no use for this, the proof is omitted.

Retracts can be useful in the theory of ordered tensor products. For instance, in Example 9.7 and Example 10.20 we proved that the projective cone does not preserve subspaces and the injective cone does not preserve quotients, but retracts are sufficiently rigid to be preserved by both (a similar role is played by complemented subspaces in the theory of normed tensor products). The following result shows that retracts can also be useful for comparing the projective and injective cones.

Proposition 12.18 (cf. [ALP19, Proposition 8])**.** *Let G and H be finite-dimensional real vector spaces, let $G_+ \subseteq G$, $H_+ \subseteq H$ be closed convex cones, and let $E \subseteq G$ and $F \subseteq H$ be retracts. If $E_+ \otimes^\pi F_+ \neq E_+ \otimes^\varepsilon F_+$, then $G_+ \otimes^\pi H_+ \neq G_+ \otimes^\varepsilon H_+$.*

Proof. We prove the contrapositive: assuming that $G_+ \otimes^\pi H_+ = G_+ \otimes^\varepsilon H_+$, we prove that $E_+ \otimes^\pi F_+ = E_+ \otimes^\varepsilon F_+$. By Remark 12.14(d), we may identify E^* with a retract of G^*. Choose positive projections $\pi_{E^*} : G^* \twoheadrightarrow E^*$ and $\pi_F : H \twoheadrightarrow F$.

Let $T : E^* \to F$ be a positive linear map. Since every positive operator $G^* \to H$ factors through a simplex cone, we may choose positive operators $S_1 : G^* \to \mathbb{R}^m$ and $S_2 : \mathbb{R}^m \to H$ (where $\mathbb{R}^m$ carries the standard cone $\mathbb{R}^m_{\geq 0}$) so that the following diagram commutes:



Now $T : E^* \to F$ factors through a simplex cone, so $T \in E_+ \otimes^\pi F_+$. ■

In Chapter 13, we will show that $G_+ \otimes^\pi H_+ \neq G_+ \otimes^\varepsilon H_+$ for a large class of cones. For combinations of standard cones, we will use Proposition 12.18 to reduce the problem to 3-dimensional cones (see Theorem 13.13). However, it was shown in [ALP19, Lemma S14] that most convex cones do not have retracts[3], so having retracts is an exceptional property. Hence, for non-standard cones G_+ and H_+, different strategies are needed to show that $G_+ \otimes^\pi H_+ \neq G_+ \otimes^\varepsilon H_+$. Several such techniques will be discussed in the next chapter.

[3]More precisely, for $n \geq 4$, the set of $(n-1)$-dimensional convex bodies whose homogenizations have an $(n-1)$-dimensional retract is meagre with respect to the Hausdorff measure.

Chapter 13

Many examples where the projective and injective cone differ

In this chapter, we provide many examples where the projective cone is closed and strictly contained in the injective cone. This proves a conjecture of Barker for nearly all closed, proper and generating convex cones in finite-dimensional spaces. Independently, the conjecture was proved in full in simultaneous work by Aubrun, Lami, Palazuelos and Plávala [ALPP21].

This chapter is based on Chapter 7 of [Dob20b].

Introduction

A question which has attracted considerable attention is under which circumstances the projective cone $E_+ \otimes^\pi F_+$ is dense in the injective cone $E_+ \otimes^\varepsilon F_+$. Birnbaum [Bir76, Prop. 3] showed that this is always the case if E and F are locally convex lattices, and gave an example which shows that it is not true in general. However, the infinite-dimensional version of this problem does not appear to be well understood.

A lot more is known in the finite-dimensional case (with closed, proper and generating cones). In this setting, the projective cone is automatically closed, so the question is now whether or not the two cones coincide. In [Bar76, p. 197], Barker asked to find precise necessary and sufficient conditions for the two cones to coincide, and later formulated the following conjecture:

Conjecture 13.1 (Barker, [Bar81, p. 277])**.** *Let E, F be finite-dimensional, and let $E_+ \subseteq E$, $F_+ \subseteq F$ be closed, proper and generating convex cones. Then $E_+ \otimes^\pi F_+ \neq E_+ \otimes^\varepsilon F_+$ unless at least one of E_+ and F_+ is a simplex cone.*

Various partial results in this direction have been obtained over the years. Before Barker formulated his conjecture, it had already been proved in the case that $F_+ = E_+^*$ by Barker and Loewy [BL75], and in the case that E_+ and F_+ are polyhedral by Poole [Poo75, Thm. 5.15]. More recently, Huber and Netzer [HN21] proved it in the case that E_+ is polyhedral and F_+ is a positive semidefinite cone (or vice versa).

In this chapter, we prove Conjecture 13.1 for nearly all[1] convex cones, thereby also providing new proofs for each of the aforementioned known cases. First, in §13.1, we give another proof for the case $F_+ = E_+^*$, and we extend the aforementioned result of Barker and Loewy [BL75] to non-proper closed cones. Then, in §13.2, we give another proof for the polyhedral case, originally due to Poole [Poo75, Thm. 5.15]. Our main contribution comes in §13.3, where we prove the conjecture when $\dim(E) \geq \dim(F)$ and E_+ is smooth or strictly convex. Finally, in §13.4, we prove it for all possible combinations of standard cones (polyhedral cones, second-order cones, and positive semidefinite cones), thereby also providing a new proof of the mixed polyhedral/positive semidefinite case settled by Huber and Netzer [HN21].

Although some of the cases we prove had been known before, all proofs in this chapter are original. However, as the manuscript [Dob20b] upon which this part is based was being written, the results from this chapter were superseded by independent work of Aubrun, Lami, Palazuelos and Plávala [ALPP21], who were able to prove Conjecture 13.1 in full generality. Our results were obtained independently around the same time, and the proofs are completely different.

13.1 The tensor product of a closed convex cone with its dual

The simplest case where the projective and injective cone are different occurs when considering the tensor product of a closed convex cone with its dual. The main result of this section, Theorem 13.2, is a slight extension of a well-known result of Barker and Loewy [BL75, Prop. 3.1] (see also [Tam77b, Thm. 4]), who proved it for convex cones that are closed, proper and generating. The proof below is much simpler than the original proof in [BL75], since it was not realized at the time that the projective tensor product of closed convex cones is automatically closed, but comparable in size with Tam's alternative proof [Tam77b, Thm. 4].

If E_+ and F_+ are closed, then by Corollary 12.2 one has $E_+^* \otimes^\pi F_+ = E_+^* \otimes^\varepsilon F_+$ if and only if every positive linear map $E \to F$ factors through a simplex cone. This language makes it much easier to think about the difference between projective and injective cones.

If T or S factors through a simplex cone, then so does the composition $S \circ T$. This shows that the separable positive operators form an *ideal* in the *semiring* of positive operators. (Ideals of operators also play an important role in the theory of normed tensor products; e.g. [DF93]. We won't make much use of this terminology.)

Theorem 13.2 (cf. [BL75, Prop. 3.1], [Tam77b, Thm. 4])**.** *Let E be finite-dimensional and let $E_+ \subseteq E$ be a closed convex cone. Then the following are equivalent:*

(i) E_+ is a simplex cone;

(ii) $\mathrm{id}_E : E \to E$ is separable (i.e. factors through a simplex cone);

(iii) for every positive linear map $T : E \to E$, one has $\mathrm{tr}(T) \geq 0$;

[1] The term 'nearly all' has a precise meaning; namely, up to a σ-porous set. Since the set of closed, proper and generating convex cones which are not smooth or not strictly convex form a σ-porous set [Zam87], the results in this chapter prove Conjecture 13.1 for nearly all E_+ and F_+.

(iv) for every finite-dimensional real vector space F and every closed convex cone $F_+ \subseteq F$, one has $E_+^* \otimes^\pi F_+ = E_+^* \otimes^\varepsilon F_+$;

(v) for every finite-dimensional real vector space F and every closed convex cone $F_+ \subseteq F$, one has $F_+ \otimes^\pi E_+ = F_+ \otimes^\varepsilon E_+$;

(vi) $E_+^* \otimes^\pi E_+ = E_+^* \otimes^\varepsilon E_+$.

Proof. (i) $\implies$ *(ii).* Let $x_1, \ldots, x_d$ be a basis of E which generates E_+, and let $x_1^*, \ldots, x_d^*$ be the corresponding dual basis. Then $x_1^*, \ldots, x_d^* \in E_+^*$, and $\mathrm{id}_E = \sum_{i=1}^d x_i^* \otimes x_i$.

(ii) $\iff$ *(iii).* The trace $\mathrm{tr} \in L(E,E)^* = (E^* \otimes E)^* = E \otimes E^*$ is the transpose of the identity $\mathrm{id}_E \in L(E,E) = E^* \otimes E$. Since $(E_+^* \otimes^\varepsilon E_+)^* = E_+ \otimes^\pi E_+^*$, we see that the trace defines a positive linear functional (i.e. $\mathrm{tr} \in (E_+^* \otimes^\varepsilon E_+)^*$) if and only if id_E is separable (i.e. $\mathrm{id}_E \in E_+^* \otimes^\pi E_+$).

(ii) $\implies$ *(iv)* and *(ii)* $\implies$ *(v).* Since $\mathrm{id}_E : E \to E$ factors through a simplex cone, so do all positive linear maps to or from E:

$$\begin{array}{ccccc} & & \mathbb{R}^m & & \\ & \nearrow & & \searrow & \\ F^* & \longrightarrow & E & \xrightarrow{\mathrm{id}_E} & E \end{array} \qquad \text{or} \qquad \begin{array}{ccccccc} & & \mathbb{R}^m & & & & \\ & \nearrow & & \searrow & & & \\ E & & & & E & \longrightarrow & F. \end{array}$$

Therefore $F_+ \otimes^\pi E_+ = F_+ \otimes^\varepsilon E_+$ and $E_+^* \otimes^\pi F_+ = E_+^* \otimes^\varepsilon F_+$.

(iv) $\implies$ *(vi)* and *(v)* $\implies$ *(vi).* Clear.

(vi) $\implies$ *(ii).* Note that $\mathrm{id}_E : E \to E$ is positive.

(ii) $\implies$ *(i).* Write $\mathrm{id}_E = \sum_{i=1}^k \varphi_i \otimes x_i$ with $\varphi_1, \ldots, \varphi_k \in E_+^*$, $x_1, \ldots, x_k \in E_+$. For arbitrary $x \in E_+$ we have $x = \mathrm{id}_E(x) = \sum_{i=1}^k \varphi_i(x) x_i$ with $\varphi_1(x), \ldots, \varphi_k(x) \geq 0$, so we see that E_+ is generated by $x_1, \ldots, x_k$. In particular, it follows that E_+ is a polyhedral cone. Furthermore, since we have $E = \mathrm{ran}(\mathrm{id}_E) \subseteq \mathrm{span}(x_1, \ldots, x_k)$, it follows that $x_1, \ldots, x_k$ must span E, so E_+ is generating. Dually, if $x, -x \in E_+$, then $\varphi_1(x) = \ldots = \varphi_k(x) = 0$, hence $x = \mathrm{id}_E(x) = \sum_{i=1}^k \varphi_i(x) x_i = 0$, which shows that E_+ is a proper cone.

Since both E_+ and E_+^* are proper polyhedral cones, each has a finite number of extremal rays generating the cone. Let $\{\psi_i\}_{i=1}^n$ and $\{y_j\}_{j=1}^m$ be (representatives of) the extremal directions of E_+^* and E_+, respectively. Writing every φ_i and every x_j as a positive combination of the extremal rays, we can expand our expression of id_E to

$$\mathrm{id}_E = \sum_{i=1}^n \sum_{j=1}^m \lambda_{ij} \psi_i \otimes y_j, \quad \text{with } \lambda_{ij} \geq 0 \text{ for all } i \text{ and all } j.$$

For every j we have $y_j = \mathrm{id}_E(y_j) = \sum_{i=1}^n \sum_{k=1}^m \lambda_{ik} \psi_i(y_j) y_k = \sum_{i=1}^n \lambda_{ij} \psi_i(y_j) y_j$, for by extremality of y_j the terms $\lambda_{ik} \psi_i(y_j) y_k$ with $k \neq j$ must be zero. It follows that $\sum_{i=1}^n \lambda_{ij} \psi_i(y_j) = 1$ for all j. Therefore:

$$\dim(E) = \mathrm{tr}(\mathrm{id}_E) = \mathrm{tr}\left(\sum_{i=1}^n \sum_{j=1}^m \lambda_{ij} \psi_i \otimes y_j \right) = \sum_{j=1}^m \sum_{i=1}^n \lambda_{ij} \psi_i(y_j) = \sum_{j=1}^m 1 = m.$$

Since $\operatorname{span}(y_1, \ldots, y_m) = \operatorname{span}(E_+) = E$, it follows that $y_1, \ldots, y_m$ is a basis of E. This proves that E_+ is a simplex cone. ■

Remark 13.3. Taking the tensor product of a space with its dual is also a common technique in the theory of normed tensor products. For instance, Theorem 13.2 is very similar to a result about the approximation property; see [DF93, Theorem 5.6].

The following corollary is immediate.

Corollary 13.4. *Let $E_+ \subseteq \mathbb{R}^n$ be a self-dual cone. Then $E_+ \otimes^\pi E_+ = E_+ \otimes^\varepsilon E_+$ if and only if E_+ is a simplex cone.*

In particular, it follows that $\mathcal{S}_+^n \otimes^\pi \mathcal{S}_+^n \neq \mathcal{S}_+^n \otimes^\varepsilon \mathcal{S}_+^n$ and $\mathcal{H}_+^n \otimes^\pi \mathcal{H}_+^n \neq \mathcal{H}_+^n \otimes^\varepsilon \mathcal{H}_+^n$ whenever $n \geq 2$. This has been known for a long time in relation to quantum theory and C^*-algebras, and is related to the difference between positive and *completely positive* operators. The interested reader is encouraged to refer to the expository article by Ando [And04, §2].

13.2 Tensor products of polyhedral cones

In this section, we prove Conjecture 13.1 for polyhedral cones (see Theorem 13.8 below). This was originally proved by Poole in 1975 [Poo75, Thm. 5.15], and rediscovered recently by Aubrun, Lami, and Palazuelos [ALP19, Result 2] and (independently) by the author [Dob20b]. We follow the proof of [Dob20b], which uses a simple combinatorial argument in terms of the face lattice.

First, we use retracts to reduce the problem to the 3-dimensional case. Using the results from §12.4 and §13.1, we can prove the following lemmas.

Lemma 13.5. *Every retract of a finite-dimensional simplex cone is a simplex cone.*

Proof. Let F_+ be a finite-dimensional simplex cone and let E_+ be a retract of F_+. By Theorem 13.2, we have $E_+^* \otimes^\pi F_+ = E_+^* \otimes^\varepsilon F_+$, so it follows from Proposition 12.18 that $E_+^* \otimes^\pi E_+ = E_+^* \otimes^\varepsilon E_+$. Another application of Theorem 13.2 shows that E_+ is a simplex cone. ■

Lemma 13.6. *Let E be finite-dimensional and let $E_+ \subseteq E$ be a proper and generating polyhedral cone. Then E_+ is a simplex cone if and only if every 3-dimensional retract of E_+ is a simplex cone.*

Proof. If $\dim(E) \leq 2$, then E_+ is automatically a simplex cone, and there are no 3-dimensional retracts, so the statement is vacuously true.

Assume $\dim(E) \geq 3$. If E_+ is a simplex cone, then every retract of E_+ is a simplex cone, by Lemma 13.5. Conversely, if E_+ is not a simplex cone, then one of the following must be true (use [Brø83, Theorem 12.19] and homogenization):

- $\dim(E_+) = 3$;
- E_+ has a facet that is not a simplex cone;

- E_+ has a vertex figure that is not a simplex cone.

Since facets and vertex figures are retracts (Proposition 12.16), it follows by induction that E_+ has a 3-dimensional retract that is not a simplex cone. ■

All that remains is to prove that the projective and injective tensor products of two 3-dimensional polyhedral cones are different, unless one of the two is a simplex cone. For this we use a combinatorial argument, based on the results from Chapter 9. (For a different proof, see [ALP19].)

The proof essentially boils down to finding a combinatorial obstruction. The high-level idea behind the proof is that, if $E_+ \otimes^\pi F_+ = E_+ \otimes^\varepsilon F_+$, then $(E_+ \otimes^\pi F_+)^* = E_+^* \otimes^\pi F_+^*$, so by Theorem 9.22 we known exactly what the extremal rays of $E_+ \otimes^\pi F_+$ and $(E_+ \otimes^\pi F_+)^*$ are. This gives us enough information to determine the face lattice of $E_+ \otimes^\pi F_+$. However, if both E_+ and F_+ have at least 4 extremal rays, then it turns out that the lattice thus obtained is not *graded* (in other words, it contains maximal chains of different lengths), which contradicts a well-known property of polyhedral cones.

The proof below uses slightly different terminology than the preceding high-level idea, and does not proceed by contradiction.

Lemma 13.7. *Let E_+ and F_+ be proper and generating polyhedral cones in $\mathbb{R}^3$. Then $E_+ \otimes^\pi F_+ = E_+ \otimes^\varepsilon F_+$ if and only if at least one of E_+ and F_+ is a simplex cone.*

Proof. A proper and generating polyhedral cone in $\mathbb{R}^3$ is the homogenization of a polygon. Let $v_1, \dots, v_m \in \mathbb{R}^3$ be (representatives of) the extremal directions of E_+ in such a way that the neighbours of v_i are v_{i-1} and v_{i+1} (modulo m). In the same way, let $w_1, \dots, w_n \in F_+$ be the extremal directions of F_+ (in cyclic order). Furthermore, let $\varphi_1, \dots, \varphi_m \in E_+^*$ and $\psi_1, \dots, \psi_n \in F_+^*$ be the extremal directions of E_+^* and F_+^*, in such a way that φ_i (resp. ψ_j) represents the facet of E_+ (resp. F_+) that contains v_i and v_{i+1} (resp. w_j and w_{j+1}).

If E_+ or F_+ is a simplex cone, then $E_+ \otimes^\pi F_+ = E_+ \otimes^\varepsilon F_+$ by Theorem 13.2. So assume that neither E_+ nor F_+ is a simplex cone, i.e. $m, n \geq 4$. We show by a combinatorial argument that $(E_+ \otimes^\pi F_+)^*$ must be larger than $E_+^* \otimes^\pi F_+^* = (E_+ \otimes^\varepsilon F_+)^*$.

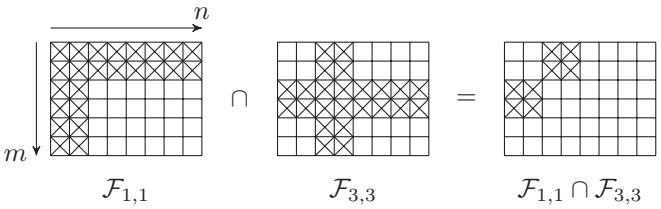
By Theorem 9.22, the extremal directions of the projective cone $E_+ \otimes^\pi F_+$ are given by $\{v_i \otimes w_j : i \in [m], j \in [n]\}$, and the extremal directions of $E_+^* \otimes^\pi F_+^*$ are given by $\{\varphi_i \otimes \psi_j : i \in [m], j \in [n]\}$. Furthermore, by Corollary 11.4(b), the extremal directions of $E_+^* \otimes^\pi F_+^*$ are also extremal for the (larger) cone $(E_+ \otimes^\pi F_+)^* = E_+^* \otimes^\varepsilon F_+^*$. To complete the proof, we show that this larger cone must have extremal directions which are not of the form $\varphi_i \otimes \psi_j$. (By Corollary 11.4(b), these must have rank ≥ 2.) Equivalently, the projective cone $E_+ \otimes^\pi F_+$ must have facets which cannot be written as the tensor product of a facet in E_+ and a facet in F_+.

Given $k \in [m]$ and $\ell \in [n]$, let $\mathcal{F}_{k,\ell}$ denote the facet of $E_+ \otimes^\pi F_+$ corresponding to the extremal direction $\varphi_k \otimes \psi_\ell \in (E_+ \otimes^\pi F_+)^* = E_+^* \otimes^\varepsilon F_+^*$. The extremal directions in $\mathcal{F}_{k,\ell}$ are those $v_i \otimes w_j$ with $i \in \{k, k+1\}$ or $j \in \{\ell, \ell+1\}$, for one has $v_i \otimes w_j \in \mathcal{F}_{k,\ell}$ if and only if $0 = \langle v_i \otimes w_j, \varphi_k \otimes \psi_\ell \rangle = \langle v_i, \varphi_k \rangle \langle w_j, \psi_\ell \rangle$. In particular, $\mathcal{F}_{k,\ell}$ contains $2m + 2n - 4$ extremal rays.

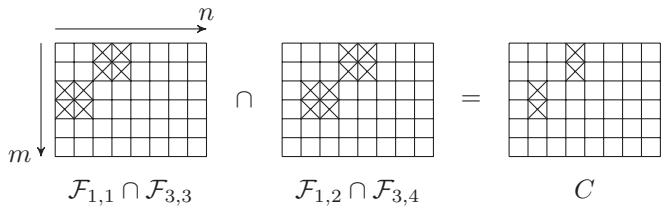
We proceed to consider a specific intersection of the $\mathcal{F}_{k,\ell}$, namely

$$C := \mathcal{F}_{1,1} \cap \mathcal{F}_{1,2} \cap \mathcal{F}_{3,3} \cap \mathcal{F}_{3,4}.$$

Clearly C is a face of $E_+ \otimes^\pi F_+$. We claim that C contains only 4 extremal rays. To that end, note that $\mathcal{F}_{1,1} \cap \mathcal{F}_{3,3}$ contains exactly 8 extremal rays, namely $v_i \otimes w_j$ with $(i,j) \in (\{1,2\} \times \{3,4\}) \cup (\{3,4\} \times \{1,2\})$. This is illustrated in the figure below (with $m = 6$ and $n = 8$).



Similarly, $\mathcal{F}_{1,2} \cap \mathcal{F}_{3,4}$ has the same pattern, but shifted one step in the second coordinate, so we see that the intersection of $\mathcal{F}_{1,1} \cap \mathcal{F}_{3,3}$ and $\mathcal{F}_{1,2} \cap \mathcal{F}_{3,4}$ contains 4 extremal rays:



(We have to be aware of a subtlety here: if $n = 4$, then the pattern of $\mathcal{F}_{1,2} \cap \mathcal{F}_{3,4}$ is "wrapped around" from right to left, but this does not affect the conclusion.)

Since C contains 4 extremal rays, we have $\dim(C) \leq 4$. However, note that the only $\mathcal{F}_{k,\ell}$ containing C are the four facets defining C. Since we know from classical polyhedral geometry that C must be contained in at least $9 - \dim(C) \geq 5$ facets, this shows that $E_+ \otimes^\pi F_+$ has facets which are not of the form $\mathcal{F}_{k,\ell}$. Equivalently, the dual cone $(E_+ \otimes^\pi F_+)^* = E_+^* \otimes^\varepsilon F_+^*$ has extremal directions which are not of the form $\varphi_i \otimes \psi_j$, so we have $E_+^* \otimes^\varepsilon F_+^* \neq E_+^* \otimes^\pi F_+^*$. By duality, it follows that $E_+ \otimes^\pi F_+ \neq E_+ \otimes^\varepsilon F_+$. ■

Theorem 13.8 ([Poo75, Thm. 5.15])**.** *Let G, H be finite-dimensional and let $G_+ \subseteq G$, $H_+ \subseteq H$ be proper and generating polyhedral cones. Then $G_+ \otimes^\pi H_+ = G_+ \otimes^\varepsilon H_+$ if and only if at least one of G_+ and H_+ is a simplex cone.*

Proof. If G_+ or H_+ is a simplex cone, then it follows from Theorem 13.2 that $G_+ \otimes^\pi H_+ = G_+ \otimes^\varepsilon H_+$. So assume that neither G_+ nor H_+ is a simplex cone. By Lemma 13.6, we may choose 3-dimensional retracts E_+ (resp. F_+) of G_+ (resp. H_+) such that neither E_+ nor F_+ is a simplex cone. It follows from Lemma 13.7 that $E_+ \otimes^\pi F_+ \neq E_+ \otimes^\varepsilon F_+$, hence it follows from Proposition 12.18 that $G_+ \otimes^\pi H_+ \neq G_+ \otimes^\varepsilon H_+$. ■

13.3 Tensor product with a smooth or strictly convex cone

In this section, we prove Conjecture 13.1 in the case that $\dim(E) \geq \dim(F)$ and E_+ is smooth or strictly convex. (This is Theorem J from Chapter 7.) The argument is based on a generalized John's decomposition of the identity.

We recall some common terminology. A *convex body* is a compact convex set $C \subseteq \mathbb{R}^n$ with non-empty interior. The (*one-sided*) *polar* of C is the set $C^\circ = \{y \in (\mathbb{R}^n)^* : \langle x, y\rangle \leq 1 \text{ for all } x \in C\}$. An *affine transformation* is an invertible affine map $\mathbb{R}^n \to \mathbb{R}^n$; that is, a map of the form $x \mapsto T_0 x + y_0$ with $T_0 \in \mathrm{GL}_n(\mathbb{R})$ and $y_0 \in \mathbb{R}^n$ fixed.

If $C_1, C_2 \subseteq \mathbb{R}^n$ are convex bodies, then a compactness argument shows that there is an affine transformation T such that $T[C_1] \subseteq C_2$ and $\mathrm{vol}(T[C_1])$ is maximal among all affine transformations T' for which $T'[C_1] \subseteq C_2$. If the maximum is attained for $T = I_n$ (the identity transformation), then we say that C_1 is *in a maximum volume position inside* C_2. Furthermore, we say that C_1 is *in John's position inside* C_2 if $C_1 \subseteq C_2$ and there exist $m \in \mathbb{N}$, $x_1, \ldots, x_m \in \partial C_1 \cap \partial C_2$, $y_1, \ldots, y_m \in \partial C_1^\circ \cap \partial C_2^\circ$ and $\lambda_1, \ldots, \lambda_m > 0$ such that $\langle x_i, y_i\rangle = 1$ for all i, and

$$I_n = \sum_{i=1}^m \lambda_i x_i \otimes y_i \qquad \text{and} \qquad 0 = \sum_{i=1}^m \lambda_i x_i = \sum_{i=1}^m \lambda_i y_i.$$

Gordon, Litvak, Meyer and Pajor [GLMP04] proved the following result, building on earlier extensions ([GPT01, BR02]) of Fritz John's classical theorem ([Joh48]).

Theorem 13.9 ([GLMP04, Theorem 3.8])**.** *Let $C_1, C_2 \subseteq \mathbb{R}^n$ be convex bodies such that C_1 is in a maximum volume position inside C_2. Then there exists $z \in \mathrm{int}(C_1)$ such that $C_1 - z$ is in John's position inside $C_2 - z$.*

For our purposes, we will only need the following (much weaker) corollary.

Corollary 13.10. *Let $C_1, C_2 \subseteq \mathbb{R}^n$ be convex bodies. Then there is an affine transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ such that $T[C_1] \subseteq C_2$ and $\partial T[C_1] \cap \partial C_2$ contains an affine basis of $\mathbb{R}^n$.*

(Equivalently, there is a T such that $T[C_1] \subseteq C_2$ and the set of points where $T[C_1]$ and C_2 touch is not contained in an affine hyperplane.)

Proof of Corollary 13.10. Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be an affine transformation such that $T[C_1]$ is in a maximum volume position inside C_2. By Theorem 13.9, we may choose $z \in \mathrm{int}(T[C_1])$, $m \in \mathbb{N}$, $x_1, \ldots, x_m \in \partial(T[C_1] - z) \cap \partial(C_2 - z)$, $y_1, \ldots, y_m \in \partial(T[C_1] - z)^\circ \cap \partial(C_2 - z)^\circ$ and $\lambda_1, \ldots, \lambda_m > 0$ such that $\langle x_i, y_i\rangle = 1$ for all i, and

$$I_n = \sum_{i=1}^m \lambda_i x_i \otimes y_i \qquad \text{and} \qquad 0 = \sum_{i=1}^m \lambda_i x_i = \sum_{i=1}^m \lambda_i y_i. \tag{$*$}$$

After an appropriate rescaling of the λ_i, the second formula in $(*)$ shows that $0 \in \mathrm{aff}(x_1, \ldots, x_m)$, so it follows that $\mathrm{aff}(x_1, \ldots, x_m) = \mathrm{span}(x_1, \ldots, x_m)$. Moreover, it follows immediately from the first formula in $(*)$ that $\mathrm{span}(x_1, \ldots, x_m) = \mathbb{R}^n$, so we

conclude that $\{x_1, \dots, x_m\}$ contains an affine basis, say $x_1, \dots, x_{n+1}$. Consequently, $x_1 + z, \dots, x_{n+1} + z$ is an affine basis in $\partial T[C_1] \cap \partial C_2$. ■

Since every closed and proper convex cone has a compact base, the following homogenization of Corollary 13.10 follows immediately.

Corollary 13.11. *Let E and F be finite-dimensional real vector spaces with $\dim(E) = \dim(F)$, and let $E_+ \subseteq E$, $F_+ \subseteq F$ be closed, proper, and generating convex cones. Then there exists a positive linear transformation $T : E \to F$ such that $\partial T[E_+] \cap \partial F_+$ contains a (linear) basis of F.*

Using the preceding results, we can prove the main result of this section.

Theorem 13.12. *Let E, F be finite-dimensional real vector spaces, and let $E_+ \subseteq E$, $F_+ \subseteq F$ be closed, proper, and generating convex cones. If $\dim(E) \geq \dim(F)$, and if E_+ is strictly convex or smooth, then one has $E_+ \otimes^\pi F_+ = E_+ \otimes^\varepsilon F_+$ if and only if F_+ is a simplex cone.*

Proof. First assume that $E_+ \otimes^\pi F_+ = E_+ \otimes^\varepsilon F_+$, with E_+ strictly convex and $\dim(E) \geq \dim(F)$. Choose an interior point $x_0 \in E_+$ and a linear subspace $G \subseteq E$ of dimension $\dim(F)$ through x_0. Then $G_+ := G \cap E_+$ is closed, proper, and generating, so by Corollary 13.11 we may choose a positive linear isomorphism $T : F^* \to G$ such that $\partial T[F_+^*] \cap \partial G_+$ contains a basis $\{b_1, \dots, b_m\}$ of G. Note that every b_i is also a boundary point of E_+ (this a basic property of topological boundaries), and therefore an extremal direction of E_+ (since E_+ is strictly convex). For all i, write $a_i := T^{-1}(b_i) \in \partial F_+^*$; then $\{a_1, \dots, a_m\}$ is a basis of F^*.

If $\iota : G \hookrightarrow E$ denotes the inclusion, then $\iota \circ T : F^* \to E$ is positive, so by assumption we may write $\iota \circ T = \sum_{i=1}^k y_i \otimes x_i$ with $x_1, \dots, x_k \in E_+$ and $y_1, \dots, y_k \in F_+$ (where the y_i act as linear functionals on F^*). Since b_i is extremal and

$$b_i = T(a_i) = \sum_{j=1}^k \langle y_j, a_i \rangle x_j,$$

it follows that at least one of the x_j is a positive multiple of b_i, and $\langle y_j, a_i \rangle = 0$ whenever x_j is not a positive multiple of b_i. In particular, if x_j is not a positive multiple of any one of the b_i, then $\langle y_j, a_i \rangle = 0$ for all i, so $y_j = 0$ (since $\{a_1, \dots, a_m\}$ is a basis of F^*). Thus, after removing the zero terms, every x_j is a positive multiple of some b_i, and so in particular belongs to $T[F_+^*]$. This shows that not only $\iota \circ T$, but also T is separable, and not only with respect to the cones F_+^* and G_+, but even with respect to the cones F_+^* and $T[F_+^*]$. Since $\mathrm{id}_{F^*} = T^{-1} \circ T$, it follows from the ideal property of separable operators that id_{F^*} is also separable. Hence, by Theorem 13.2, F_+^* is a simplex cone. Since F_+ is closed, it follows that $F_+ = F_+^{**}$ is also a simplex cone.

Now assume that $E_+ \otimes^\pi F_+ = E_+ \otimes^\varepsilon F_+$ with E_+ smooth and $\dim(E) \geq \dim(F)$. By duality (see Corollary 12.13(b)), it follows that

$$E_+^* \otimes^\pi F_+^* = (E_+ \otimes^\varepsilon F_+)^* = (E_+ \otimes^\pi F_+)^* = E_+^* \otimes^\varepsilon F_+^*.$$

Since E_+ is smooth, the dual cone E_+^* is strictly convex, so it follows from the first part of the proof that F_+^* must be a simplex cone. Since F_+ is closed, it follows that $F_+ = F_+^{**}$ is also a simplex cone. ■

13.4 Tensor products of standard cones; applications to operator systems

In this section, we prove Conjecture 13.1 for all combinations of standard cones, thereby proving Theorem K and Corollary M from Chapter 7. Just as in §13.2, we use retracts (see §12.4) to reduce the problem to the three-dimensional case.

Standard cones

By combining the results obtained thus far, we can easily prove Conjecture 13.1 for all combinations of standard cones.

Theorem 13.13 (cf. [Poo75, HN21])**.** *Let G, H be finite-dimensional real vector spaces, and let $G_+ \subseteq G$, $H_+ \subseteq H$ be closed, proper, and generating convex cones. Assume that each of G_+ and H_+ is one of the following (all combinations allowed):*

(i) a polyhedral cone;

(ii) a second-order cone $\mathcal{L}^n$;

(iii) a (real or complex) positive semidefinite cone $\mathcal{S}_+^n$ or $\mathcal{H}_+^n$.

Then one has $G_+ \otimes^\pi H_+ = G_+ \otimes^\varepsilon H_+$ if and only if at least one of G_+ and H_+ is a simplex cone.

Proof. Suppose that neither G_+ nor H_+ is a simplex cone. We claim that G_+ (resp. H_+) has a three-dimensional retract E_+ (resp. F_+) which is isomorphic with one of the following:

- the three-dimensional Lorentz cone $\mathcal{L}^3$ (which is isomorphic to $\mathcal{S}_+^2$);
- a proper and generating polyhedral cone $P \subseteq \mathbb{R}^3$ that is not a simplex cone.

To prove the claim, we distinguish three cases:

- If G_+ is polyhedral, then this follows from Lemma 13.6.
- If $G_+ = \mathcal{L}^n$, then the assumption that G_+ is not a simplex cone forces $n \geq 3$. Hence it follows from Example 12.15(d) that $\mathcal{L}^3$ is a retract of G_+.
- If $G_+ = \mathcal{S}_+^n$ or $\mathcal{H}_+^n$, then the assumption that G_+ is not a simplex cone forces $n \geq 2$, so it follows from Example 12.15(c) and Example 12.15(f) that $\mathcal{S}_+^2$ ($\cong \mathcal{L}^3$) is a retract of G_+.

Next, we show that $E_+ \otimes^\pi F_+ \neq E_+ \otimes^\varepsilon F_+$, again distinguishing three cases.

- If $E_+ = F_+ = \mathcal{L}^3$, then this follows from Corollary 13.4, since the Lorentz cone is self-dual.
- If $E_+ = \mathcal{L}^3$ and F_+ is polyhedral (or vice versa), then this follows from Theorem 13.12, since $\mathcal{L}^3$ is strictly convex and $\dim(E) = \dim(F)$.
- The case where both E_+ and F_+ are polyhedral follows from Lemma 13.7.

Since E_+ and F_+ are retracts of G_+ and H_+ satisfying $E_+ \otimes^\pi F_+ \neq E_+ \otimes^\varepsilon F_+$, it follows from Proposition 12.18 that $G_+ \otimes^\pi H_+ \neq G_+ \otimes^\varepsilon H_+$. ■

Operator systems

Some of our results can be reformulated in terms of operator systems. Let $C \subseteq \mathbb{R}^d$ be a closed, proper, and generating convex cone, and let $n \in \mathbb{N}_1$ be a positive integer. Following notation from [FNT17], we denote the projective and injective tensor products $\mathcal{H}_+^n \otimes^\pi C$ and $\mathcal{H}_+^n \otimes^\varepsilon C$ by $C_n^{\min}$ and $C_n^{\max}$, respectively, and we write $C^{\min} = \{C_n^{\min}\}_{n=1}^\infty$ and $C^{\max} = \{C_n^{\max}\}_{n=1}^\infty$.

Corollary 13.14 (Special case of [ALPP21, Corollary 2])**.** *Let $C \subseteq \mathbb{R}^d$ be a closed, proper, and generating convex cone. If $d \leq 4$, or if C is strictly convex, or smooth, or polyhedral, or (real or complex) positive semidefinite, then the following are equivalent:*

(i) C is a simplex cone;

(ii) the minimal and maximal operator systems $C^{\min}$ and $C^{\max}$ are equal;

(iii) there exists $n \geq 2$ for which $C_n^{\min} = C_n^{\max}$;

(iv) one has $C_2^{\min} = C_2^{\max}$.

Proof. (i) $\implies$ (ii). This follows from Theorem 13.2.

(ii) $\implies$ (iii). Trivial.

(iii) $\implies$ (iv). If $\mathcal{H}_+^n \otimes^\pi C = \mathcal{H}_+^n \otimes^\varepsilon C$ for some $n \geq 2$, then it follows from Proposition 12.18 that $\mathcal{H}_+^2 \otimes^\pi C = \mathcal{H}_+^2 \otimes^\varepsilon C$, since $\mathcal{H}_+^2$ is a retract of $\mathcal{H}_+^n$, by Example 12.15(c).

(iv) $\implies$ (i). First we prove that $\mathcal{H}_+^2$ is strictly convex. Indeed, the interior points of $\mathcal{H}_+^2$ are the positive definite matrices, so the boundary points are the singular 2×2 positive semidefinite matrices. Consequently, a non-zero boundary point of $\mathcal{H}_+^2$ must be a rank one positive semidefinite matrix, which is known to be extremal (it is a positive multiple of a rank one orthogonal projection).

Now suppose that C is of one of the forms described in the theorem, but not a simplex cone. If $d \leq \dim(\mathcal{H}^2) = 4$, or if C is smooth or strictly convex, then it follows from Theorem 13.12 that $\mathcal{H}_+^2 \otimes^\pi C \neq \mathcal{H}_+^2 \otimes^\varepsilon C$. If C is positive semidefinite or polyhedral (but not a simplex cone), then it follows from Theorem 13.13 that $\mathcal{H}_+^2 \otimes^\pi C \neq \mathcal{H}_+^2 \otimes^\varepsilon C$. Either way, we have $C_2^{\min} \neq C_2^{\max}$. ■

13.5 Closing remarks

As mentioned before, Aubrun, Lami, Palazuelos and Plávala [ALPP21] independently proved Conjecture 13.1 in full generality.

Theorem 13.15 ([ALPP21, Theorem A])**.** *Let E, F be finite-dimensional real vector spaces, and let $E_+ \subseteq E$, $F_+ \subseteq F$ be closed, proper, and generating convex cones. Then one has $E_+ \otimes^\pi F_+ = E_+ \otimes^\varepsilon F_+$ if and only if at least one of E_+ and F_+ is a simplex cone.*

The following example shows that this is no longer true if we omit the requirement that E_+ or F_+ is proper or generating.

Example 13.16. Let $F_+ \subseteq F$ be a "partial simplex cone"; that is, a cone generated by $m < \dim(F)$ linearly independent vectors $x_1, \dots, x_m \in F$. Furthermore, let E be another finite-dimensional space, and let $E_+ \subseteq E$ be an arbitrary closed, proper, and generating cone.

Since E_+ is generating, every positive linear map $T : E \to F$ has its range contained in $\operatorname{span}(F_+)$. Since $\operatorname{span}(F_+)$ is ordered by a simplex cone, this shows that every positive linear map $E \to F$ is simplex-factorable, hence $E_+^* \otimes^\pi F_+ = E_+^* \otimes^\varepsilon F_+$. △

In the preceding example, E_+^* and F_+ are closed and proper and E_+^* is generating, so the requirement that F_+ is generating cannot be omitted from Theorem 13.15. Furthermore, by duality, we also have $E_+ \otimes^\pi F_+^* = E_+ \otimes^\varepsilon F_+^*$, which shows that the requirement that F_+ is proper cannot be omitted either.

In a sense, a partial simplex cone (or its dual) is almost a simplex cone. In fact, we can extend Theorem 13.15 to show that all examples must be of this form. If E_+ is a closed convex cone, then we define the *proper reduction* $\operatorname{prop}(E_+)$ of E_+ as the positive cone of $\operatorname{span}(E_+)/\operatorname{lin}(E_+)$. Equivalently, choose subspaces $E_1, E_2, E_3 \subseteq E$ such that $E_1 = \operatorname{lin}(E_+)$, $E_1 \oplus E_2 = \operatorname{span}(E_+)$, and $E_1 \oplus E_2 \oplus E_3 = E$; then the proper reduction of E_+ is the positive cone $(E_2)_+ := E_2 \cap E_+$ of E_2, viewed as a closed, proper, and generating cone in E_2. It is readily verified that the projection $E \to E_2$, $(e_1, e_2, e_3) \mapsto e_2$ is positive (every projection onto $\operatorname{span}(E_+)$ is positive, and adding or subtracting elements of the lineality space does not affect positivity), so $\operatorname{prop}(E_+)$ is a retract of E_+. Therefore Theorem 13.15 has the following extension.

Corollary 13.17. *Let E, F be finite-dimensional real vector spaces, and let $E_+ \subseteq E$, $F_+ \subseteq F$ be closed convex cones. If $E_+ \otimes^\pi F_+ = E_+ \otimes^\varepsilon F_+$, then at least one of* $\operatorname{prop}(E_+)$ *and* $\operatorname{prop}(F_+)$ *is a simplex cone.*

Proof. Since $\operatorname{prop}(E_+)$ and $\operatorname{prop}(F_+)$ are retracts of E_+ and F_+, it follows from Proposition 12.18 that $\operatorname{prop}(E_+) \otimes^\pi \operatorname{prop}(F_+) = \operatorname{prop}(E_+) \otimes^\varepsilon \operatorname{prop}(F_+)$. But $\operatorname{prop}(E_+)$ and $\operatorname{prop}(F_+)$ are closed, proper, and generating, so it follows from Theorem 13.15 that at least one of $\operatorname{prop}(E_+)$ and $\operatorname{prop}(F_+)$ must be a simplex cone. ■

The converse is not true; it can happen that $\operatorname{prop}(E_+)$ and $\operatorname{prop}(F_+)$ are simplex cones but $E_+ \otimes^\pi F_+ \neq E_+ \otimes^\varepsilon F_+$. This is because $\operatorname{prop}(E_+) \otimes^\pi \operatorname{prop}(F_+) =$

$\mathrm{prop}(E_+) \otimes^\varepsilon \mathrm{prop}(F_+)$ does not necessarily imply $E_+ \otimes^\pi F_+ = E_+ \otimes^\varepsilon F_+$; the implication of Proposition 12.18 only runs in the other direction. As an extreme example, consider the case where $E_+ = \{0\}$ and $F_+ = F$; then one has $E_+ \otimes^\pi F_+ = \{0\}$ but $E_+ \otimes^\varepsilon F_+ = E \otimes F$. More generally, Theorem 13.2 shows that $E_+^* \otimes^\pi E_+ \neq E_+^* \otimes^\varepsilon E_+$ whenever E_+ is not proper or not generating, regardless of whether or not $\mathrm{prop}(E_+)$ is a simplex cone.

Chapter 14

Open problems for Part III

We conclude Part III with a few open problems.

This chapter is based on Chapter 8 of [Dob20b].

Does the closure of the projective cone preserve higher faces? We showed in §9.4 that the projective cone preserves faces. Furthermore, it follows from Proposition 11.7 that the closure of the projective cone preserves extremal rays (provided that E_+ and F_+ are weakly closed). However, we suspect that this result is of limited use in practice, because infinite-dimensional cones often do not have sufficiently many extremal rays. Does the closure of the projective cone also preserve higher faces, in a sense similar to Theorem 9.13? In particular, if E_+ and F_+ are weakly closed proper cones and $M \subseteq E_+$ and $M \subseteq F_+$ are faces, then is $\overline{M \otimes^\pi N}^w$ a face of $\overline{E_+ \otimes^\pi F_+}^w$?

As a partial result, it follows from Proposition 11.7(a) that $(\operatorname{span}(M) \otimes \operatorname{span}(N)) \cap \overline{E_+ \otimes^\pi F_+}^w$ is a face of $\overline{E_+ \otimes^\pi F_+}^w$, but this can in principle be larger than $\overline{M \otimes^\pi N}^w$.

Does the projective norm preserve extreme points? We showed in §9.6 that the algebraic tensor product $\operatorname{conv}(C \otimes_s D)$ of symmetric convex sets C and D preserves proper faces. Is this still true if we pass to the closure of $\operatorname{conv}(C \otimes_s D)$? In particular, we do not known whether the projective norm preserves extreme points of the closed unit ball. See also Remark 9.32.

Is the projective cone still semisimple in the completed projective tensor product? We showed in §11.3 that every reasonable crosscone in the algebraic tensor product $E \otimes F$ is semisimple whenever the base cones E_+ and F_+ are semisimple. This is no longer true if we pass to the completed locally convex tensor product $E \tilde{\otimes}_\alpha F$, because the injective cone $E_+ \tilde{\otimes}^\varepsilon_\alpha F_+$ is not proper if the natural map $E \tilde{\otimes}_\alpha F \to E \tilde{\otimes}_\varepsilon F$ fails to be injective (see Corollary 10.23). However, we don't know the answer if we match the projective cone with the projective norm; see Question 11.16.

Is there a way to determine all extremal rays of the injective cone? In §10.6, we showed that the injective cone preserves extremal rays. Corollary 11.4(b) shows that all extremal rays of tensor rank one are of this form, but Example 10.51

shows that there may be extremal rays of higher rank. Is there a way to determine all extremal rays of the injective cone?

We note that it was already pointed out earlier by Tam [Tam92, p. 75] that this appears to be a difficult problem. In fact, it is already difficult for proper and generating polyhedral cones in finite-dimensional spaces; see [BCG13]. Sufficient conditions for some $z \in E \otimes F$ to be an extremal direction of the injective cone were studied by various authors in the context of positive operators; see for instance [Tam95] and the references contained therein.

Are there other interesting tensor cones? In this paper, we have drawn parallels between normed and ordered tensor products. Taking this a step further, we may define a *tensor cone* as a way of choosing for each pair of preordered vector spaces (E, E_+) and (F, F_+) a reasonable crosscone $E_+ \otimes^\alpha F_+$ in the tensor product $E \otimes F$, in such a way that positive linear maps are preserved; that is, $(T \otimes S)[E_+ \otimes^\alpha F_+] \subseteq G_+ \otimes^\alpha H_+$ whenever $T[E_+] \subseteq G_+$ and $S[F_+] \subseteq H_+$. The projective/injective cone defines a tensor cone which behaves similarly to its normed counterpart. Many more tensor *norms* are known, but to our knowledge no other tensor *cones* have been studied in the literature.[1] Are there other interesting and/or natural tensor cones?

Every tensor cone defines a tensor norm via a construction similar to Proposition 9.25. Conversely, can every tensor norm be extended to a tensor cone in a natural way? Are there cone-theoretic analogues of Grothendieck's 14 natural tensor norms?

An important difference between cones and norms is that there is no notion of two cones being equivalent. So even if this programme would succeed, the resulting theory might not be as nice as the normed theory. It is unclear if a cone-theoretic analogue of, say, Grothendieck's inequality, can exist. Therefore it is also conceivable that there are more than 14 natural tensor cones.

Our findings about faces and extremal rays of the injective cone already show that ordered tensor products are not completely analogous to normed tensor products (see Remark 10.52), so perhaps there are limits to the analogy.

Is there a full proof of Theorem 13.15 without relying on computer algebra? In [ALPP21], Aubrun, Lami, Palazuelos and Plávala proved a very general result about the difference between the projective and injective cone (stated as Theorem 13.15 above), which contains all our results from Chapter 13 as a special case. Their proof uses an ingenious geometric argument, but one step in their proof relies on a couple of large computations, which the authors verified using computer algebra. On the other hand, the special cases that we proved in this paper only rely on relatively simple geometric arguments. Can we find an alternative proof of Theorem 13.15 which does not rely on computer algebra?

Is there an infinite-dimensional version of Theorem 13.15? The theorem of Aubrun, Lami, Palazuelos and Plávala (stated as Theorem 13.15 above) provides precise

[1] A few other cones have been defined in the tensor product of specific types of ordered vector spaces (e.g. for Archimedean Riesz spaces [Fre72], or for Archimdean ordered vector spaces [GL88, GK10]), but not for arbitrary ordered vector spaces, and rarely in connection with a positive mapping property.

necessary and sufficient conditions for the projective cone to be equal to the injective cone, when the base cones are closed, proper and generating in finite-dimensional spaces. Can this be extended to infinite-dimensional spaces? Are there examples where the projective cone $E_+ \otimes^\pi F_+$ is dense in the injective cone $E_+ \otimes^\varepsilon F_+$ (with respect to some compatible topology on $E \otimes F$) where E_+ and F_+ are weakly closed but neither is a lattice cone? How about in the *completed* locally convex tensor product?

Apart from Theorem 13.15, very few results in this direction are known. A small positive result is [Bir76, Prop. 3], which shows that $E_+ \otimes^\pi F_+$ is dense in $E_+ \otimes^\varepsilon F_+$ with respect to the projective topology whenever E and F are locally convex lattices.

Appendix **A**

Ideals, faces, and duality

This appendix discusses the basic properties of faces and ideals in preordered vector spaces. Although many of these results are known in some form, the connections between these concepts are not particularly well-known. The main body of Part III of this dissertation draws heavily on these connections, especially on the results from §A.1.

This chapter is based on Appendix A of [Dob20b].

Introduction

Certain special subsets of a convex set, the so-called *faces*, play an important role in convex geometry. For instance, convex polytopes are commonly studied in terms of their face lattice, and the extreme points and extremal rays of convex sets play an important role in convex analysis and optimization. The face structure of a convex cone has also been studied extensively; see for instance [SW70, §2.13] and the works of Barker and Tam [Bar73, Bar78a, Tam85, Tam92].

Similarly, certain special subspaces, the so-called *order ideals*, play a special role in the theory of ordered vector spaces. In [Kad51], Kadison used maximal order ideals in the proof of his celebrated representation theorem, and Bonsall continued the study of order ideals in [Bon54]. However, as attention shifted from general ordered vector spaces to lattice-ordered (i.e. Riesz) spaces, order ideals appear to have been forgotten in favour of lattice ideals (sometimes also called order ideals). As a result, the theory of order ideals is not so well-known.

In this appendix, we develop/recall the basics of order ideals in a general preordered vector space. In §A.1, we give several different equivalent definitions of an order ideal, and we show that the order ideals of a preordered vector space (E, E_+) are closely related to the faces of the positive cone E_+. This is very useful, as it allows us to quotient out a face, which is one of the main tools in the construction of faces of the projective cone in §9.4.

In §A.2, we outline the homomorphism and isomorphism theorems for ideals in ordered vector spaces. As an application, we show that the maximal order ideals are precisely the supporting hyperplanes of the positive cone. This shows that general (non-maximal) order ideals can be thought of as being the "supporting subspaces" of the positive cone.

Finally, in §A.3, we extend the theory of dual faces (see [Bar78a, Tam85]) to cones in infinite-dimensional spaces. We show that it is now necessary to make a distinction between *dual* and *exposed* faces, although the two notions coincide if the ambient space is a separable normed space.

A.1 Faces and ideals

Let E be a preordered vector space with positive cone $E_+ \subseteq E$, and let $\leq$ be the vector preorder corresponding with E_+. A subset $M \subseteq E$ is *full* (or *order-convex*) if $x \leq y \leq z$ with $x, z \in M$ implies $y \in M$. A non-empty subset $M \subseteq E_+$ that is a convex cone in its own right is called a *subcone*. Recall that a *face* (or *extremal set*) of E_+ is a (possibly empty) convex subset $M \subseteq E_+$ such that, if M intersects the relative interior of a line segment in E_+, then M contains both endpoints of that segment.

Proposition A.1. *A non-empty subset $M \subseteq E_+$ is a face if and only if it is a full subcone.*

Proof. "$\Longrightarrow$". Suppose that M is a face. First we show that M is a convex cone. If $x \in M$ and $\lambda > 1$, then the line segment from 0 to λx contains x in its relative interior, so the endpoints 0 and λx must also belong to M. Then, since M is convex, for all $\lambda \in [0, 1]$ we also have $\lambda x \in M$. Since a face is convex by assumption, we conclude that M is a convex cone.

To see that M is full, suppose that $x \leq y \leq z$ with $x, z \in M$. Then $x, z \in E_+$, so in particular we have $y \geq x \geq 0$, or in other words, $y \in E_+$. Furthermore, we have $z - y \in E_+$ (since $y \leq z$), so it follows that $y + 2(z - y) \in E_+$. Since $z = y + (z - y)$ is in the relative interior of the line segment from y to $y + 2(z - y)$, we must have $y \in M$, which proves that M is full.

"$\Longleftarrow$". Suppose that $M \subseteq E_+$ is a full subcone, and suppose that $x, z \in E_+$ and $\lambda \in (0, 1)$ are such that $y := \lambda x + (1 - \lambda)z$ belongs to M. If $x = z$, then evidently $x = z = y \in M$, so assume $x \neq z$. Then y lies in the relative interior of the line segment between x and z, so for small enough $\mu < 0$ the point $\mu x + (1 - \mu)y$ also lies on this line segment. In particular, $\mu x + (1 - \mu)y \geq 0$, or equivalently, $y \geq \frac{-\mu}{1-\mu} x$. But we have $\frac{-\mu}{1-\mu} > 0$ and $x \geq 0$, so we find $0 \leq \frac{-\mu}{1-\mu} x \leq y$. Since M is full, it follows that $\frac{-\mu}{1-\mu} x \in M$, and therefore $x \in M$. Analogously, $z \in M$. ∎

Note that the lineality space $\operatorname{lin}(E_+) = E_+ \cap -E_+ = \{x \in E_+ : 0 \leq x \leq 0\}$ and the cone E_+ itself are full subcones, and therefore faces of E_+. Furthermore, clearly every face contains $\operatorname{lin}(E_+)$ and is contained in E_+, so these are the unique minimal and maximal faces of E_+.[1]

Next we come to the subject of ideals. If $I \subseteq E$ is a subspace, then we define the *quotient cone* $(E/I)_+$ to be the image of E_+ under the canonical map $E \to E/I$.

Proposition A.2 (Equivalent definitions of an order ideal)**.** *Let (E, E_+) be a preordered vector space. For a linear subspace $I \subseteq E$, the following are equivalent:*

[1] In order-theoretic terms, these are the *least* and the *greatest* element in the set of faces (ordered by inclusion).

(i) I *is full;*

(ii) if $-y \leq x \leq y$ *and* $y \in I$*, then* $x \in I$*;*

(iii) if $0 \leq x \leq y$ *and* $y \in I$*, then* $x \in I$*;*

(iv) $I_+ := I \cap E_+$ *is a face of* E_+*;*

(v) the quotient cone $(E/I)_+$ *is proper.*

Proof. *(i)* $\Longrightarrow$ *(ii).* Clear.

(ii) $\Longrightarrow$ *(iii).* Suppose that $0 \leq x \leq y$ and $y \in I$. Then we also have $-y \leq 0 \leq x$, so we find $-y \leq x \leq y$. It follows that $x \in I$.

(iii) $\Longrightarrow$ *(iv).* Every linear subspace is a convex cone, and the intersection of two convex cones is a convex cone, so $I_+ \subseteq E_+$ is a subcone. If $x \leq y \leq z$ with $x, z \in I_+$, then in particular $0 \leq y \leq z$ with $z \in I$, so we have $y \in I$. Furthermore, we have $y \geq x \geq 0$, so $y \in I_+$, which shows that I_+ is full. By Proposition A.1, I_+ is a face of E_+.

(iv) $\Longrightarrow$ *(v).* Let $z \in (E/I)_+ \cap -(E/I)_+$ be given, then we may choose $x, y \in E_+$ such that $z = \pi(x) = \pi(-y)$. It follows that $\pi(x + y) = 0$, so $x + y \in I$. As such, we have $0 \leq x \leq x + y$ and $0 \leq y \leq x + y$ with $0, x + y \in I_+$, so we find $x, y \in I_+$ (since I_+ is full). It follows that $z = 0$, which shows that $(E/I)_+$ is a proper cone.

(v) $\Longrightarrow$ *(i).* Clearly the natural map $\pi : E \to E/I$ is positive. Suppose that $x \leq y \leq z$ with $x, z \in I$, then $0 = \pi(x) \leq \pi(y) \leq \pi(z) = 0$, so it follows that $\pi(y) = 0$ (since $(E/I)_+$ is a proper cone). Therefore: $y \in I$. ■

A subspace I satisfying any one (and therefore all) of the conditions of Proposition A.2 is called an *order ideal*, or simply *ideal* if no ambiguity can arise (i.e. if the space does not have additional algebraic structure). Order ideals have been studied since the 1950s (e.g. [Kad51, Bon54]), but the link between ideals and faces does not appear to be well-known.

We give a few useful ways to obtain ideals or faces:

Proposition A.3. *Let* E, F *be vector spaces and let* $E_+ \subseteq E$*,* $F_+ \subseteq F$ *be convex cones.*

(a) If $M \subseteq E_+$ *is a non-empty face, then* $\operatorname{span}(M)$ *is an ideal satisfying* $M = \operatorname{span}(M) \cap E_+$*.*

(b) If $T : E \to F$ *is a positive linear map and if* $J \subseteq F$ *is an ideal, then* $T^{-1}[J] \subseteq E$ *is an ideal.*

Proof.

(a) Clearly $M \subseteq \operatorname{span}(M) \cap E_+$. Moreover, since M is a convex cone, every $x \in \operatorname{span}(M)$ can be written as $x = m_1 - m_2$ with $m_1, m_2 \in M$. If furthermore $x \in E_+$, then we find $0 \leq x \leq m_1$ (because $m_1 - x = m_2 \geq 0$), and therefore $x \in M$ (because M is full). This shows that $M = \operatorname{span}(M) \cap E_+$. It follows from Proposition A.2(iv) that $\operatorname{span}(M)$ is an ideal.

(b) If $x \le y \le z$ and $x, z \in T^{-1}[J]$, then $T(x) \le T(y) \le T(z)$ with $T(x), T(z) \in J$. Since J is full, it follows at once that $T(y) \in J$, which shows that $T^{-1}[J]$ is also full. ■

It follows from Proposition A.2(iv) and Proposition A.3(a) that the map $I \mapsto I_+$ defines a surjective many-to-one correspondence between the ideals and the non-empty faces.

A first (and rather important) application of this correspondence is given in Proposition A.4(b) below. If $\varphi : E \to \mathbb{R}$ is a positive linear functional, then $\ker(\varphi) \cap E_+$ is easily seen to be a face, and faces of this type are called *exposed*. This can be generalized in the following way: if F is any vector space with a proper cone $F_+ \subseteq F$, and if $T : E \to F$ is a positive linear map, then it is still relatively easy to see that $\ker(T) \cap E_+$ is a face. (It is crucial that F_+ is proper!) Although not every face is exposed, the following result shows that this slight extension already captures all faces.

Proposition A.4. *Let E be a vector space and let $E_+ \subseteq E$ be a convex cone.*

(a) (cf. [Bon54, §2, p. 403]) A subspace $I \subseteq E$ is an ideal if and only if it occurs as the kernel of a positive linear map $T : E \to F$ with F_+ proper.

(b) A non-empty subset $M \subseteq E_+$ is a face if and only if it can be written as $M = \ker(T) \cap E_+$ with $T : E \to F$ positive and F_+ proper.

Proof.

(a) If $I \subseteq E$ is an ideal, then $(E/I)_+$ is a proper cone (by Proposition A.2(v)), the map $T : E \to E/I$ is positive, and $I = \ker(T)$.

Conversely, if $T : E \to F$ is a positive linear map with $F_+ \subseteq F$ a proper cone, then $\{0\} \subseteq F_+$ is an ideal (because F_+ is proper), so it follows from Proposition A.3(b) that $\ker(T)$ is an ideal in E.

(b) If $M \subseteq E_+$ is a face, then $I := \operatorname{span}(M)$ is an ideal with $M = I \cap E_+$ (by Proposition A.3(a)), so $(E/I)_+$ is a proper cone, the map $T : E \to E/I$ is positive, and $M = \ker(T) \cap E_+$.

Conversely, if $T : E \to F$ is a positive linear map with $F_+ \subseteq F$ a proper cone, then it follows from (a) that $\ker(T)$ is an ideal, so $\ker(T) \cap E_+$ is a face. ■

Remark A.5. Just as $\operatorname{lin}(E_+)$ and E_+ are the smallest and the largest face of E_+, the smallest and the largest ideals of E are $\operatorname{lin}(E_+)$ and E. Apart from this, the maximal ideals $\neq E$ are of some interest; see Corollary A.12 below.

For now, we show that the smallest ideal has the following special property.

Proposition A.6. *Let E be a vector space, $E_+ \subseteq E$ a convex cone, and $I \subseteq E$ a subspace. Then the quotient $\pi_I : E \to E/I$ is bipositive if and only if $I \subseteq \operatorname{lin}(E_+)$.*

In particular, the only ideal $I \subseteq E$ for which the quotient $\pi_I : E \to E/I$ is bipositive is the minimal ideal $I = \operatorname{lin}(E_+)$.

Proof. Bipositivity of the quotient $E \to E/I$ means that, if $x \in E_+$ and $x + I = y + I$, then $y \in E_+$. Equivalently: if $x \in E_+$ and $z \in I$, then $x + z \in E_+$. Evidently this is the case if and only if $I \subseteq E_+$ (use that $0 \in E_+$). But I is a subspace, so we have $I \subseteq E_+$ if and only if $I \subseteq \mathrm{lin}(E_+)$.

If I is an ideal, then we have $\mathrm{lin}(E_+) \subseteq I$ (every ideal contains the minimal ideal), so the second conclusion follows immediately. ■

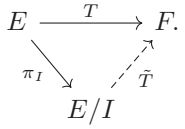
Remark A.7. If E_+ is proper and if F_+ is arbitrary, then every bipositive map $T : E \to F$ is automatically injective, since $\ker(T) = T^{-1}[\{0\}] \subseteq T^{-1}[F_+] = E_+$ is a subspace contained in E_+, which must therefore be $\{0\}$. The preceding proposition shows that this is no longer true if E_+ is not proper.

A.2 The homomorphism and isomorphism theorems

In connection with the ideal theory, we investigate to which extent the homomorphism and isomorphism theorems hold for ordered vector spaces.

The homomorphism theorem and the third isomorphism theorem hold true for ordered vector spaces.

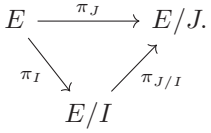
Proposition A.8 (Homomorphism theorem)**.** *Let E, F be vector spaces, $E_+ \subseteq E$, $F_+ \subseteq F$ convex cones, $T : E \to F$ a positive linear map, and $I \subseteq E$ a subspace with $I \subseteq \ker(T)$. Then there is a unique positive linear map $\tilde{T} : E/I \to F$ for which the following diagram commutes:*



Proof. Since $I \subseteq \ker(T)$, there is a unique linear map $\tilde{T} : E/I \to F$ for which the diagram commutes. This map is automatically positive: if $y \in (E/I)_+$, then there is some $x \in E_+$ such that $y = \pi_I(x)$, and it follows that $\tilde{T}(y) = T(x) \in T[E_+] \subseteq F_+$. ■

Proposition A.9 (Third isomorphism theorem)**.** *Let E be a vector space, $E_+ \subseteq E$ a convex cone, and $I \subseteq J \subseteq E$ subspaces. Then the natural isomorphism $(E/I)/(J/I) \cong E/J$ is bipositive for the respective quotient cones. Furthermore, the bijective correspondence $J \mapsto J/I$ between the subspaces $I \subseteq J \subseteq E$ and the subspaces of E/I restricts to a bijective correspondence of order ideals (in other words, J is an ideal in E if and only if J/I is an ideal in E/I).*

Proof. We have the following commutative diagram of linear maps:



To see that the natural isomorphism $(E/I)/(J/I) \cong E/J$ is bipositive, note that pushforwards commute: an element of E/J belongs to either one of the pushforward cones $(E/J)_+$ and $((E/I)/(J/I))_+$ if and only if it has a positive element of E in its preimage.

Since a subspace is an ideal if and only if the quotient cone is proper, it follows immediately that J is an ideal in E if and only if J/I is an ideal in E/I. ■

Analogous results hold for *closed* ideals in ordered *topological* vector spaces. (We assume no compatibility between the positive cone and the topology, so questions of continuity and positivity are completely separate from one another.)

Contrary to the preceding results, the first and second isomorphism theorems fail for ordered vector spaces. We only have the following weaker statements, of which the (simple) proofs are omitted.

Proposition A.10 (Partial first isomorphism theorem)**.** *Let E, F be vector spaces, $E_+ \subseteq E$, $F_+ \subseteq F$ convex cones, and $T : E \to F$ a positive linear map. Then the natural linear isomorphism $E/\ker(T) \xrightarrow{\sim} \operatorname{ran}(T)$ is positive, but not necessarily bipositive.*

Counterexample against bipositivity: $E = F$ and $T = \mathrm{id}_E$, but E_+ strictly contained in F_+.

Proposition A.11 (Partial second isomorphism theorem)**.** *Let F be a vector space, $F_+ \subseteq F$ a convex cone, $E \subseteq F$ a subspace, and $I \subseteq F$ an order ideal. Then $E + I$ is a subspace of F, $E \cap I$ is an order ideal of E, and the natural linear isomorphism $E/(E \cap I) \xrightarrow{\sim} (E+I)/I$, $x+(E\cap I) \mapsto x+I$ is positive, but not necessarily bipositive.*

Counterexample against bipositivity: $F = \mathbb{R}^2$ with standard cone, and $E, I \subseteq F$ two different one-dimensional subspaces, each of which meets F_+ only in 0. Then $E_+ := E \cap F_+ = \{0\}$, so the cone of $E/(E \cap I)$ is $\{0\}$, whereas the cone of $(E+I)/I = \mathbb{R}^2/I$ is generating.

Classification of maximal order ideals

As an application of the preceding results, we show that the third isomorphism theorem gives a geometric characterization of the maximal order ideals.

Following common terminology from algebra, we say that an order ideal $I \subseteq E$ is *proper* if $I \neq E$, and *maximal* if it is proper and not contained in another proper ideal. Furthermore, we say that a preordered vector space E is *simple* if E_+ is proper and if E has exactly two order ideals (namely, the trivial ideals $\{0\}$ and E). Bonsall [Bon54, Theorem 2] proved that an ordered vector space is simple if and only if it is one-dimensional (with either the standard cone or the zero cone).[2] Combining this with Proposition A.9, we find:

Corollary A.12. *The maximal order ideals of E are precisely the supporting hyperplanes of E_+.*

[2]Bonsall also includes $\{0\}$ among the simple ordered spaces, but we require *exactly* two ideals. (Similarly, we believe that 1 is not prime, the empty topological space is not connected, etc.) This is just a matter of convention.

Proof. It is easy to see that the supporting hyperplanes of E_+ are precisely the kernels of the non-zero positive linear functionals. (For a proof, see e.g. [Dob20a, Proposition 4.1].) Furthermore, it follows from Proposition A.9 that an ideal $I \subseteq E$ is maximal if and only if E/I is simple.

If $\varphi : E \to \mathbb{R}$ is a non-zero positive linear functional, then $\ker(\varphi)$ is an ideal, which is maximal since $E/\ker(\varphi)$ is one-dimensional and therefore simple.

Conversely, if $I \subseteq E$ is a maximal ideal, then E/I is simple, so $\dim(E/I) = 1$ and the quotient cone $(E/I)_+$ is either $\{0\}$ or isomorphic to the standard cone $\mathbb{R}_{\geq 0}$. Either way, we can choose a linear isomorphism $E/I \xrightarrow{\sim} \mathbb{R}$ which is positive (but not necessarily bipositive), so that the composition $\varphi : E \to E/I \to \mathbb{R}$ is a non-zero positive linear functional with $I = \ker(\varphi)$. ■

For more on maximal ideals, see [Bon54, §4].

A.3 Dual and exposed faces

In the finite-dimensional setting (with closed cones), dual faces are well studied in the literature (see e.g. [Bar78a, Wei12]). We outline a theory of face duality in dual pairs.

A *positive pairing* is a dual pair $\langle E, F\rangle$ of (real) preordered vector spaces such that $\langle x, y\rangle \geq 0$ whenever $x \in E_+$, $y \in F_+$. In this case we say $\langle E_+, F_+\rangle$ is a *positive pair.*[3]

Given a positive pair $\langle E_+, F_+\rangle$ and a non-empty subset $N \subseteq F_+$, we define the (*pre*)*dual face*

$$^{\diamond}N := E_+ \cap {}^{\perp}N = \big\{x \in E_+ \,:\, \langle x, y\rangle = 0 \text{ for all } y \in N\big\}.$$

Analogously, for a non-empty subset $M \subseteq E_+$ we define the *dual face*

$$M^{\diamond} := F_+ \cap M^{\perp} = \big\{y \in F_+ \,:\, \langle x, y\rangle = 0 \text{ for all } x \in M\big\}.$$

Note that $^{\diamond}N$ (resp. $M^{\diamond}$) depends not only on N (resp. M), but also implicitly on E_+ (resp. F_+).

Proposition A.13. *Let $\langle E_+, F_+\rangle$ be a positive pair.*

(a) If $N \subseteq F_+$ is non-empty, then $^{\diamond}N$ is a face of E_+.

(b) If $N_1 \subseteq N_2 \subseteq F_+$ are non-empty, then $^{\diamond}N_1 \supseteq {}^{\diamond}N_2$.

(c) If $N \subseteq F_+$ is non-empty, then $N \subseteq (^{\diamond}N)^{\diamond}$ and $^{\diamond}N = {}^{\diamond}((^{\diamond}N)^{\diamond})$.

Similar statements hold with N and $^{\diamond}N$ replaced by M and $M^{\diamond}$.

Proof.

(a) Every $y \in F_+$ defines a positive linear functional $\varphi_y : E \to \mathbb{R}$, $x \mapsto \langle x, y\rangle$. As such, the set $E_+ \cap \ker(\varphi_y)$ is a face, by Proposition A.3(b). Since we can write

$$^{\diamond}N = \bigcap_{y \in N} (E_+ \cap \ker(\varphi_y)),$$

[3] There is a slight abuse of notation here, for if E_+ and F_+ are not generating, then the positive pair depends not only on E_+ and F_+, but also on E and F (but this will cause no confusion).

it follows that ${}^{\diamond}N$ is also a face of E_+.

(b) This follows from the definition, since ${}^{\perp}N_1 \supseteq {}^{\perp}N_2$.

(c) If $y \in N$, then by definition one has $\langle x, y\rangle = 0$ for all $x \in {}^{\diamond}N$, so it follows that $y \in ({}^{\diamond}N)^{\diamond}$. This proves the inclusion $N \subseteq ({}^{\diamond}N)^{\diamond}$.

Write $M := {}^{\diamond}N$. It follows from the preceding argument that $M \subseteq {}^{\diamond}(M^{\diamond}) = {}^{\diamond}(({}^{\diamond}N)^{\diamond})$. On the other hand, combining the inclusion $N \subseteq ({}^{\diamond}N)^{\diamond}$ with (b), we find $M = {}^{\diamond}N \supseteq {}^{\diamond}(({}^{\diamond}N)^{\diamond})$, so we conclude that equality holds. ■

A face $M \subseteq E_+$ is said to be an $\langle E_+, F_+\rangle$-*dual face* if $M = {}^{\diamond}N$ for some non-empty subset $N \subseteq F_+$, and an $\langle E_+, F_+\rangle$-*exposed face* if there is some $y_0 \in F_+$ such that $M = E_+ \cap \ker(\varphi_{y_0})$, or equivalently, if M is the $\langle E_+, F_+\rangle$-dual face of a singleton. Likewise, the faces $N \subseteq F_+$ of the form $N = M^{\diamond}$ (resp. $N = \{x_0\}^{\diamond}$) are the $\langle F_+, E_+\rangle$-dual (resp. $\langle F_+, E_+\rangle$-exposed) faces of F_+.

The operations $M \mapsto M^{\diamond}$ and $N \mapsto {}^{\diamond}N$ define a so-called *Galois connection* (see e.g. [Ber15, §6.5] for the definition). It follows that the set of $\langle E_+, F_+\rangle$-dual faces, ordered by inclusion, forms a complete lattice, which we denote as $\mathscr{F}_{\langle E_+, F_+\rangle}$.

If $\langle E_+, G_+\rangle$ is a positive pair and if $F \subseteq G$ and $F_+ \subseteq F \cap G_+$, then evidently one has $\mathscr{F}_{\langle E_+, F_+\rangle} \subseteq \mathscr{F}_{\langle E_+, G_+\rangle}$, but the inclusion $\mathscr{F}_{\langle E_+, F_+\rangle} \hookrightarrow \mathscr{F}_{\langle E_+, G_+\rangle}$ should not be expected to be a lattice homomorphism.

Given a dual pair $\langle E, E'\rangle$ and a convex cone $E_+ \subseteq E$, the most natural lattice of dual faces in E_+ is the lattice $\mathscr{F}_{\langle E_+, E'_+\rangle}$, where $E'_+ \subseteq E'$ is the dual cone of E_+. (This is the largest of all lattices $\mathscr{F}_{\langle E_+, F_+\rangle}$ with $F_+ \subseteq E'$.) The $\langle E_+, E'_+\rangle$-dual (resp. $\langle E_+, E'_+\rangle$-exposed) faces will simply be called the *dual* (resp. *exposed*) faces of E_+.

The difference between dual and exposed faces

If E is finite-dimensional and if E_+ is closed, then every dual face is exposed, so $\mathscr{F}_{\langle E_+, E^*_+\rangle}$ is simply the lattice of exposed faces (see e.g. [Bar78a]). We intend to show that things become more complicated in the infinite-dimensional case. We illustrate these subtleties by establishing various equivalent definitions of dual and exposed faces.

For notational simplicity, we formulate the results in the remainder of this appendix not for dual pairs but for locally convex spaces. We recall some basic theory. If $\mathfrak{T}$ is a locally convex topology on E that is compatible with the dual pair $\langle E, E'\rangle$, then a subspace $I \subseteq E$ is $\mathfrak{T}$-closed if and only if it is weakly closed. If this is the case, then the quotient E/I is once again a (Hausdorff) locally convex space, and $(E/I)' \cong I^{\perp}$ as vector spaces. Furthermore, if $\mathfrak{T}$ is the weak topology $\sigma(E, E')$, then E/I carries the weak topology $\sigma(E/I, I^{\perp})$ (see e.g. [Con07, §V.2] or [Köt83, §22]).

We shall say that a convex cone $E_+ \subseteq E$ is *quasi-semisimple* if E'_+ separates points on E_+; that is, if for every $x \in E_+$ there is some $\varphi_x \in E'_+$ such that $\langle x, \varphi_x\rangle > 0$. This is equivalent to the (geometric) requirement that $E_+ \cap \operatorname{lin}(\overline{E_+}) = \{0\}$, since $\operatorname{lin}(\overline{E_+}) = {}^{\perp}(E'_+)$. It follows that a quasi-semisimple cone is automatically proper. Clearly every semisimple cone (in particular, every closed proper cone in a locally convex space) is quasi-semisimple.

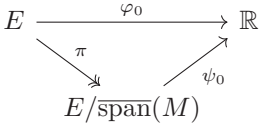
Proposition A.14. *Let E be locally convex. For a face $M \subseteq E_+$ the following are equivalent:*

(i) M is exposed.

(ii) There exists some $\varphi_0 \in M^\diamond$ such that for all $x \in E_+ \setminus M$ one has $\langle x, \varphi_0 \rangle > 0$.

(iii) $M = E_+ \cap \overline{\operatorname{span}}(M)$, and the quotient $(E/\overline{\operatorname{span}}(M))_+$ admits a strictly positive continuous linear functional.

Proof. (i) $\Longrightarrow$ (iii). Choose $\varphi_0 \in E'_+$ such that $M = E_+ \cap \ker(\varphi_0)$. Then $\overline{\operatorname{span}}(M) \subseteq \ker(\varphi_0)$, so it follows that $M = E_+ \cap \overline{\operatorname{span}}(M)$ and that φ_0 factors through $E/\overline{\operatorname{span}}(M)$:



If $E/\overline{\operatorname{span}}(M)$ is equipped with the quotient cone, then $\psi_0 : E/\overline{\operatorname{span}}(M) \to \mathbb{R}$ is strictly positive.

(iii) $\Longrightarrow$ (ii). We have $(E/\overline{\operatorname{span}}(M))' \cong \overline{\operatorname{span}}(M)^\perp = M^\perp$. Consequently, if $\psi_0 : E/\overline{\operatorname{span}}(M) \to \mathbb{R}$ is continuous and strictly positive, then the composition $\varphi_0 : E \xrightarrow{\pi} E/\overline{\operatorname{span}}(M) \xrightarrow{\psi_0} \mathbb{R}$ is continuous and positive, and belongs to $M^\perp$. It follows that $\varphi_0 \in E'_+ \cap M^\perp = M^\diamond$. Furthermore, every $x \in E_+ \setminus M$ satisfies $\langle x, \varphi_0 \rangle = \langle \pi x, \psi_0 \rangle > 0$, since ψ_0 is strictly positive.

(ii) $\Longrightarrow$ (i). The requirement $\{\varphi_0\} \subseteq M^\diamond$ ensures that $M \subseteq {}^\diamond(M^\diamond) \subseteq {}^\diamond\{\varphi_0\}$, and the assumption that $\langle x, \varphi_0 \rangle > 0$ for all $x \in E_+ \setminus M$ guarantees that ${}^\diamond\{\varphi_0\} \subseteq M$. ■

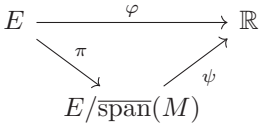
Proposition A.15. *Let E be locally convex. For a face $M \subseteq E_+$ the following are equivalent:*

(i) M is a dual face.

(ii) For every $x \in E_+ \setminus M$ there is some $\varphi_x \in M^\diamond$ such that $\langle x, \varphi_x \rangle > 0$.

(iii) $M = E_+ \cap \overline{\operatorname{span}}(M)$, and the quotient $(E/\overline{\operatorname{span}}(M))_+$ is quasi-semisimple.

Proof. (i) $\Longrightarrow$ (iii). Choose some non-empty $N \subseteq E'_+$ such that $M = {}^\diamond N = E_+ \cap {}^\perp N$. Then $\overline{\operatorname{span}}(M) \subseteq {}^\perp N$, so it follows that $M = E_+ \cap \overline{\operatorname{span}}(M)$ and that every $\varphi \in N$ factors through $E/\overline{\operatorname{span}}(M)$:



Write $\mathcal{K}$ for the positive cone of $E/\overline{\operatorname{span}}(M)$, and let $y \in \mathcal{K}$ be such that $\langle y, \psi \rangle = 0$ for all $\psi \in \mathcal{K}'$. Choose $x \in E_+$ such that $y = \pi(x)$, then for every $\varphi \in N$ we may choose

some $\psi \in \mathcal{K}'$ such that $\varphi = \psi \circ \pi$, and therefore $\langle x, \varphi\rangle = \langle y, \psi\rangle = 0$. It follows that $x \in M$, and therefore $y = 0$, showing that $\mathcal{K}$ is quasi-semisimple.

(iii) $\Longrightarrow$ *(ii).* Let $x \in E_+ \setminus M$ be given. Then x is mapped to a non-zero positive vector in $E/\overline{\operatorname{span}}(M)$, so there is a positive continuous linear functional $\psi_x : E/\overline{\operatorname{span}}(M) \to \mathbb{R}$ such that $\langle \pi x, \psi_x\rangle > 0$ (by quasi-semisimplicity). Now the composition $\varphi_x : E \xrightarrow{\pi} E/\overline{\operatorname{span}}(M) \xrightarrow{\psi_x} \mathbb{R}$ is continuous and positive, and $\langle x, \varphi_x\rangle = \langle \pi x, \psi_x\rangle > 0$.

(ii) $\Longrightarrow$ *(i).* We have $M \subseteq {}^\diamond(M^\diamond)$, and the assumption that every $x \in E_+ \setminus M$ admits some $\varphi_x \in M^\diamond$ such that $\langle x, \varphi_x\rangle > 0$ guarantees that ${}^\diamond(M^\diamond) \subseteq M$. ■

The subtle difference between exposed and dual faces becomes apparent by comparing Proposition A.14(ii) with Proposition A.15(ii): the only difference is the order of the quantifiers!

In the finite-dimensional setting, it is well-known that the dual faces are precisely the exposed faces (see e.g. [Bar78a]). We extend this to separable normed spaces. Counterexamples in other settings will be given below.

Theorem A.16 (Compare [Sch60, Proposition 15.2])**.** *Let E be locally convex, and let $M \subseteq E_+$ be a face of the form $E_+ \cap I$, where $I \subseteq E$ is a closed subspace. If E/I admits a separable norm compatible with the dual pair $\langle E/I, (E/I)'\rangle$ ($= \langle E/I, I^\perp\rangle$), then M is a dual face if and only if M is exposed.*

Proof. Every exposed face is dual. For the converse, suppose that M is a dual face, and let $\|\cdot\|$ be a separable norm compatible with the dual pair $\langle E/I, I^\perp\rangle$. We shall understand E/I and $(E/I)'$ to be equipped with the respective norm topologies. Since E/I is separable, its dual $(E/I)'$ and every subset thereof is weak-$*$ separable (this is because it can be written as the union of a countable family of separable metrizable spaces; see [Köt83, §21.3.(5)]). As such, we may choose a weak-$*$ dense countable subset $N = \{\varphi_k\}_{k=1}^\infty$ in the dual cone $(E/I)'_+$. Define $\psi := \sum_{k=1}^\infty \frac{\varphi_k}{2^k\|\varphi_k\|}$; this is well-defined because $(E/I)'$ is a Banach space.[4] Since $(E/I)'_+$ is a closed convex cone, we have $\psi \in (E/I)'_+$.

We claim that ψ is a strictly positive functional. To that end, let $x \in (E/I)_+$ be such that $\psi(x) = 0$. For all k we have $\varphi_k(x) \geq 0$, but $\psi(x) = \sum_{k=1}^\infty \frac{\varphi_k(x)}{2^k\|\varphi_k\|} = 0$, so we must have $\varphi_k(x) = 0$. It follows that $x \in {}^\perp N = {}^\perp((E/I)'_+) = \operatorname{lin}(\overline{(E/I)_+})$. Since M is a dual face, the quotient face $(E/I)_+$ is quasi-semisimple, so $(E/I)_+ \cap \operatorname{lin}(\overline{(E/I)_+}) = \{0\}$. It follows that $x = 0$, which shows that ψ is a strictly positive functional. We conclude that M is exposed. ■

Corollary A.17. *A face of finite codimension is dual if and only if it is exposed.*

Corollary A.18. *In a separable normed space, the dual faces are precisely the exposed faces.*

Corollary A.19. *If E is a separable normed space and if E_+ is closed, then $\operatorname{lin}(E_+)$ is exposed.*

[4] Technically this is not entirely well-defined; if $\varphi_k = 0$, then we must replace $\frac{\varphi_k}{2^k\|\varphi_k\|}$ by 0.

After all, $\operatorname{lin}(E_+) = \operatorname{lin}(\overline{E_+}) = {}^{\perp}(E'_+) = {}^{\diamond}(E'_+)$ is a dual face.

Corollary A.20. *Every quasi-semisimple cone (in particular, every closed proper cone) in a separable normed space admits a strictly positive continuous linear functional.*

In general, not every dual face is exposed. As a generic example, let E_+ be a cone that is semisimple but does not admit a strictly positive functional. Then $\{0\}$ is a dual face, for by semisimplicity, E'_+ separates points, so ${}^{\diamond}(E'_+) = {}^{\perp}(E'_+) = \{0\}$. However, $\{0\}$ is not exposed, since there is no strictly positive functional.

We give two concrete realizations of this generic example: one in an inseparable Hilbert space, and one in a separable Fréchet space. These examples show that the preceding corollaries cannot easily be extended beyond the setting of separable normed spaces.

Example A.21. Let Ω be an uncountable set, and consider the Hilbert space $E = \ell^2_{\mathbb{R}}(\Omega)$ with the non-negative cone $E_+ = \{x \in \ell^2_{\mathbb{R}}(\Omega) : x_\omega \geq 0 \text{ for all } \omega \in \Omega\}$. Then E_+ is semisimple, so $\{0\}$ is a dual face. However, E_+ does not admit a strictly positive functional, since every vector in $E' = \ell^2_{\mathbb{R}}(\Omega)$ is zero in all but at most countably many coordinates. △

Example A.22. Let s be the space of all (real) sequences with the topology of pointwise convergence. Then s is a separable Fréchet space with topological dual $s' = c_{00}$, the space of sequences of finite support. (The last statement is a special case of duality between products and locally convex direct sums; see [Köt83, §22.5.(2)].) The non-negative cone in s is closed and proper, so $\{0\}$ is a dual face. However, there is no strictly positive functional. △

As a final remark, we point out that certain faces stand no chance of being either exposed or a dual face. We know from Proposition A.3(a) that every face is the positive part of an ideal, but a dual or exposed face must be the positive part of a *closed* ideal. It may happen that $E_+ \cap \overline{\operatorname{span}}(M)$ is larger than M, in which case Proposition A.15(iii) (resp. Proposition A.14(iii)) shows that M cannot be a dual (resp. exposed) face.

As a concrete example, let $E = \ell^\infty_{\mathbb{R}}$ with its usual cone and norm, and let $M = E_+ \cap c_{00}$ be the set of all non-negative sequences with finite support. Then M is a face, but $E_+ \cap \overline{\operatorname{span}}(M) = E_+ \cap c_0$ is the (larger) set of all non-negative sequences converging to 0. Therefore M is not exposed or a dual face.

Glossary of notation (Part III)

$\mathrm{Bil}(E \times F)$	the space of all bilinear forms $E \times F \to \mathbb{R}$	107
$\mathscr{Bil}(E \times F)$	the space of continuous bilinear forms $E \times F \to \mathbb{R}$	107
$\mathfrak{Bil}(E \times F)$	the space of separately continuous bilinear forms $E \times F \to \mathbb{R}$	107
$b(M, \cdot\,)$	the set $\{b(x, \cdot\,) : x \in M\}$	138
$b(\cdot\,, N)$	the set $\{b(\cdot\,, y) : y \in N\}$	138
$b(x_0, \cdot\,)$	the linear functional $y \mapsto b(x_0, y)$	107
$b(\cdot\,, y_0)$	the linear functional $x \mapsto b(x, y_0)$	107
$\mathscr{C}(C)$	the homogenization of C	126
C°	the one-sided polar of C	179
E^*	the algebraic dual space of E	105
E'	the topological dual space of E	105
$\langle E, E' \rangle$	dual pair	106
$E \,\tilde{\otimes}_\alpha\, F$	the completion of $E \otimes F$ with respect to a compatible locally convex topology α on $E \otimes F$	109
$E \circledast F$	the space of separately weak-$*$ continuous bilinear forms $E' \times F' \to \mathbb{R}$	130
E_+	the positive cone of a preordered vector space E	110
$\overline{E_+}^{\,w}$	the weak closure of E_+ (= the bipolar cone with respect to $\langle E, E' \rangle$)	110
E_+^*	the algebraic dual cone of E_+	110, 164
E_+'	the topological dual cone of E_+	110
$E_+ \otimes^\varepsilon F_+$	the injective cone in the algebraic tensor product $E \otimes F$	93, 94, 129
$E_+ \,\tilde{\otimes}^\varepsilon_\alpha\, F_+$	the injective cone in the completed locally convex tensor product $E \,\tilde{\otimes}_\alpha\, F$	94, 129
$E_+ \otimes^\pi F_+$	the projective cone in the algebraic tensor product $E \otimes F$	93, 94, 115
$E_+ \,\tilde{\otimes}^\pi_\alpha\, F_+$	the projective cone in the completed locally convex tensor product $E \,\tilde{\otimes}_\alpha\, F$	115
E_w	the space E equipped with the $\sigma(E, E')$-topology	106
E'_{w*}	the space E' equipped with the $\sigma(E', E)$-topology	106

$\mathcal{H}^n$	the real vector space of complex $n \times n$ hermitian matrices	164
$\mathcal{H}^n_+$	the cone of complex $n \times n$ positive semidefinite matrices	164
$I \circledcirc J$	the upper ideal of $E \circledast F$ defined by the ideals I and J	144
$I \circledwedge J$	the lower ideal of $E \circledast F$ defined by the ideals I and J	144
$\mathrm{L}(E, F)$	the space of linear maps $E \to F$	106
$\mathfrak{L}(E, F)$	the space of continuous linear maps $E \to F$	106
$\mathrm{lin}(E_+)$	the lineality space $\mathrm{lin}(E_+) = E_+ \cap -E_+$ of the convex cone E_+	93, 110, 164
$\mathcal{L}^n$	the second-order (= Lorentz) cone $\mathcal{L}^n \subseteq \mathbb{R}^n$	164
$\overline{M}^{\,w}$	the weak closure of the set M	106
$M^\perp$	the set $\{\varphi \in E' : \langle x, \varphi\rangle = 0 \text{ for all } x \in M\}$	
$M^\diamond$	the dual face of M	195
$M \otimes_s N$	the entry-wise tensor product of the sets M and N	107
$M \circledcirc^\varepsilon N$	the upper face of $(E \circledast F)_+$ defined by the faces M and N	139
$M \circledcirc^\pi N$	the upper face of $E_+ \otimes^\pi F_+$ defined by the faces M and N	120
$M \circledwedge^\varepsilon N$	the lower face of $(E \circledast F)_+$ defined by the faces M and N	139
$M \circledwedge^\pi N$	the lower face of $E_+ \otimes^\pi F_+$ defined by the faces M and N	120
$M' \ltimes N$	the set $\{b \in E \circledast F : b(M', \cdot) \subseteq N\}$	138
$M \rtimes N'$	the set $\{b \in E \circledast F : b(\cdot, N') \subseteq M\}$	138
$\overline{N}^{\,w*}$	the weak-$*$ closure of the set N	106
${}^\perp N$	the set $\{x \in E : \langle x, \varphi\rangle = 0 \text{ for all } \varphi \in N\}$	
${}^\diamond N$	the predual face of N	195
$\mathrm{rext}(E_+)$	the set of extremal directions of E_+	113
$\sigma(E, E')$	the weak topology on E	106
$\sigma(E', E)$	the weak-$*$ topology on E'	106
$\mathcal{S}^n$	the space of real $n \times n$ symmetric matrices	164
$\mathcal{S}^n_+$	the cone of real $n \times n$ positive semidefinite matrices	164
T^*	the algebraic adjoint $F^* \to E^*$ of a linear map $T : E \to F$	106
T'	the topological adjoint $F' \to E'$ of a continuous linear map $T : E \to F$	106
$T \otimes S$	the linear map $E \otimes F \to G \otimes H$ determined by the linear maps $T : E \to G$ and $S : F \to H$	116, 133
$T \circledast S$	the linear map $E \circledast F \to G \circledast H$ determined by the linear maps $T \in \mathfrak{L}(E_w, G_w)$ and $S \in \mathfrak{L}(F_w, H_w)$	133
$T \boxtimes S$	the linear map $\mathrm{Bil}(E' \times F') \to \mathrm{Bil}(G' \times H')$ determined by the linear maps $T \in \mathfrak{L}(E_w, G_w)$ and $S \in \mathfrak{L}(F_w, H_w)$	133

Index (Part III)

Bibliography

[ABJ18] Y. Azouzi, M. A. Ben Amor, and J. Jaber. The tensor product of f-algebras. *Quaestiones Mathematicae*, **41**(3):359–369, 2018. doi:10.2989/16073606.2017.1382018.

[AC13] O. Amini and L. Caporaso. Riemann–Roch theory for weigthed graphs and tropical curves. *Advances in Mathematics*, **240**:1–23, 2013. doi:10.1016/j.aim.2013.03.003.

[ALP19] G. Aubrun, L. Lami, and C. Palazuelos. Universal entangleability of non-classical theories, 2019. Preprint. URL: https://arxiv.org/abs/1910.04745v1.

[ALPP21] G. Aubrun, L. Lami, C. Palazuelos, and M. Plávala. Entangleability of cones. *Geometric and Functional Analysis*, **31**(2):181–205, 2021. doi:10.1007/s00039-021-00565-5.

[AM10] O. Amini and M. Manjunath. Riemann–Roch for sub-lattices of the root lattice A_n. *Electronic Journal of Combinatorics*, **17**:#R124, 2010. doi:10.37236/396.

[AM20] I. Aidun and R. Morrison. On the gonality of Cartesian products of graphs. *Electronic Journal of Combinatorics*, **27**(4):#P4.52, 2020. doi:10.37236/9307.

[And04] T. Ando. Cones and norms in the tensor product of matrix spaces. *Linear Algebra and its Applications*, **379**:3–41, 2004.

[AR18] S. Atanasov and D. Ranganathan. A note on Brill–Noether existence for graphs of low genus. *Michigan Mathematical Journal*, **67**(1):175–198, 2018. doi:10.1307/mmj/1519095622.

[AS17] G. Aubrun and S.J. Szarek. *Alice and Bob Meet Banach*, volume 223 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2017.

[ASU13] N. Alon, A. Shpilka, and C. Umans. On sunflowers and matrix multiplication. *Comput. Complexity*, **22**(2):219–243, 2013. doi:10.1007/s00037-013-0060-1.

[AT07] C.D. Aliprantis and R. Tourky. *Cones and Duality*, volume 84 of *Graduate Studies in Mathematics*. American Mathematical Society, 2007.

[Bac17] S. Backman. Riemann–Roch theory for graph orientations. *Advances in Mathematics*, **309**:655–691, 2017. doi:10.1016/j.aim.2017.01.005.

[Bak08] M. Baker. Specialization of linear systems from curves to graphs. *Algebra & Number Theory*, **2**(6):613–653, 2008.

[Bar73] G.P. Barker. The lattice of faces of a finite dimensional cone. *Linear Algebra and its Applications*, **7**(1):71–82, 1973.

[Bar76] G.P. Barker. Monotone norms and tensor products. *Linear and Multilinear Algebra*, **4**(3):191–199, 1976.

[Bar78a] G.P. Barker. Faces and duality in convex cones. *Linear and Multilinear Algebra*, **6**(3):161–169, 1978.

[Bar78b] G.P. Barker. Perfect cones. *Linear Algebra and its Applications*, **22**:211–221, 1978.

[Bar81] G.P. Barker. Theory of cones. *Linear Algebra and its Applications*, **39**:263–291, 1981.

[Bau96] H. Bauer. *Probability Theory*, volume 23 of *De Gruyter Studies in Mathematics*. De Gruyter, Berlin, New York, 1996.

[BB72] U. Bertelè and F. Brioschi. *Nonserial Dynamic Programming*, volume 91 of *Mathematics in Science and Engineering*. Academic Press, 1972.

[BB82] T.C. Brown and J.P. Buhler. A density version of a geometric Ramsey theorem. *Journal of Combinatorial Theory, Series A*, **32**(1):20–34, 1982. doi:10.1016/0097-3165(82)90062-0.

[BCC+17] J. Blasiak, T. Church, H. Cohn, J.A. Grochow, E. Naslund, W.F. Sawin, and C. Umans. On cap sets and the group-theoretic approach to matrix multiplication. *Discrete Analysis*, **2017:3**, 2017. 27pp. doi:10.19086/da.1245.

[BCG13] T. Bogart, M. Contois, and J. Gubeladze. Hom-polytopes. *Mathematische Zeitschrift*, **273**(3–4):1267–1296, 2013.

[BCW22a] H.L. Bodlaender, G. Cornelissen, and M. van der Wegen. Problems hard for treewidth but easy for stable gonality, 2022. Preprint. URL: https://arxiv.org/abs/2202.06838.

[BCW22b] H.L. Bodlaender, G. Cornelissen, and M. van der Wegen. Problems hard for treewidth but easy for stable gonality. In M.A. Bekos and M. Kaufmann, editors, *Graph-Theoretic Concepts in Computer Science, WG 2022*, volume 13453 of *Lecture Notes in Computer Science*, pages 84–97. Springer, Cham, 2022. doi:10.1007/978-3-031-15914-5_7.

[BDGS20] H.L. Bodlaender, J. van Dobben de Bruyn, D. Gijswijt, and H. Smit. Constructing tree decompositions of graphs with bounded gonality. In D. Kim, R.N. Uma, Z. Cai, and D.H. Lee, editors, *Computing and Combinatorics, COCOON 2020*, volume 12273 of *Lecture Notes in Computer Science*, pages 384–396. Springer, Cham, 2020. doi:10.1007/978-3-030-58150-3_31.

[BDGS22] H.L. Bodlaender, J. van Dobben de Bruyn, D. Gijswijt, and H. Smit. Constructing tree decompositions of graphs with bounded gonality. *Journal of Combinatorial Optimization*, **44**(4):2681–2699, 2022. doi:10.1007/s10878-021-00762-w.

[Ber15] G.M. Bergman. *An Invitation to General Algebra and Universal Constructions*. Universitext. Springer, second edition, 2015.

[BGY22] M.A. Ben Amor, Ö. Gok, and D. Yaman. Tensor products of ideals and projection bands, 2022. Preprint. URL: http://arxiv.org/abs/2207.13796.

[BHN97] R. Bacher, P. de la Harpe, and T. Nagnibeda. The lattice of integral flows and the lattice of integral cuts on a finite graph. *Bull. Soc. math. France*, **125**(2):167–198, 1997.

[BIL+21] M. Bläser, C. Ikenmeyer, V. Lysikov, A. Pandey, and F.O. Schreyer. On the orbit closure containment problem and slice rank of tensors. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2565–2584, 2021. doi:10.1137/1.9781611976465.152.

[Bir76] D.A. Birnbaum. Cones in the tensor product of locally convex lattices. *American Journal of Mathematics*, **98**(4):1049–1058, 1976.

[BJ16] M. Baker and D. Jensen. Degeneration of linear series from the tropical point of view and applications. In M. Baker and S. Payne, editors, *Nonarchimedean and tropical geometry*, Simons Symposia, pages 365–433. Springer, 2016. doi:10.1007/978-3-319-30945-3_11.

[BK12] M. Bateman and N.H. Katz. New bounds on cap sets. *Journal of the American Mathematical Society*, **25**(2):585–613, 2012. doi:10.1090/S0894-0347-2011-00725-X.

[BL75] G.P. Barker and R. Loewy. The structure of cones of matrices. *Linear Algebra and its Applications*, **12**(1):87–94, 1975.

[Bla16] A. Blanco. On the positive approximation property. *Positivity*, **20**(3):719–742, 2016.

[BLP87] G.P. Barker, M. Laidacker, and G. Poole. Projectionally exposed cones. *SIAM Journal on Algebraic Discrete Methods*, **8**(1):100–105, 1987.

[BN07] M. Baker and S. Norine. Riemann–Roch and Abel–Jacobi theory on a finite graph. *Advances in Mathematics*, **215**(2):766–788, 2007. doi:10.1016/j.aim.2007.04.012.

[Bod96] H.L. Bodlaender. A linear-time algorithm for finding tree-decompositions of small treewidth. *SIAM Journal on Computing*, **25**(6):1305–1317, 1996. doi:10.1137/S0097539793251219.

[Bod98] H.L. Bodlaender. A partial k-arboretum of graphs with bounded treewidth. *Theoretical Compututer Science*, **209**(1-2):1–45, 1998. doi:10.1016/S0304-3975(97)00228-4.

[Bon54] F.F. Bonsall. Sublinear functionals and ideals in partially ordered vector spaces. *Proceedings of the London Mathematical Society, 3rd series*, **4**(1):402–418, 1954.

[Bou90] J. Bourgain. On arithmetic progressions in sums of sets of integers. In A. Baker, B. Bollobás, and A. Hajnal, editors, *A tribute to Paul Erdős*, pages 105–110. Cambridge University Press, 1990. doi:10.1017/CBO9780511983917.008.

[BR02] J. Bastero and M. Romance. John's decomposition of the identity in the non-convex case. *Positivity*, **6**(1):1–16, 2002.

[Brø83] A. Brøndsted. *An Introduction to Convex Polytopes*, volume 90 of *Graduate Texts in Mathematics*. Springer, 1983.

[BS20] T.F. Bloom and O. Sisask. Breaking the logarithmic barrier in Roth's theorem on arithmetic progressions, 2020. Preprint. URL: https://arxiv.org/abs/2007.03528.

[BT22] G. Buskes and P. Thorn. Two results on Fremlin's Archimedean Riesz space tensor product, 2022. Preprint. URL: http://arxiv.org/abs/2206.06283.

[BWZ21] H.L. Bodlaender, M. van der Wegen, and T.C. van der Zanden. Stable divisorial gonality is in NP. *Theory of Computing Systems*, **65**(2):428–440, 2021. doi:10.1007/s00224-020-10019-4.

[Cap12] L. Caporaso. Algebraic and combinatorial Brill–Noether theory. In V. Alexeev, A. Gibney, E. Izadi, J. Kollár, and E. Looijenga, editors, *Compact moduli spaces and vector bundles*, Contemporary Mathematics, pages 69–85. American Mathematical Society, 2012. doi:10.1090/conm/564/11150.

[CDPR12] F. Cools, J. Draisma, S. Payne, and E. Robeva. A tropical proof of the Brill–Noether theorem. *Advances in Mathematics*, **230**(2):759–776, 2012. doi:10.1016/j.aim.2012.02.019.

[CFK+15] M. Cygan, F.V. Fomin, Ł. Kowalik, D. Lokshtanov, D. Marx, M. Pilipczuk, M. Pilipczuk, and S. Saurabh. *Parameterized Algorithms*. Springer, 2015. doi:10.1007/978-3-319-21275-3.

[CKK15] G. Cornelissen, F. Kato, and J. Kool. A combinatorial Li–Yau inequality and rational points on curves. *Mathematische Annalen*, **361**(1-2):211–258, 2015. doi:10.1007/s00208-014-1067-x.

[CLM15] L. Caporaso, Y. Len, and M. Melo. Algebraic and combinatorial rank of divisors on finite graphs. *Journal de Mathématiques Pures et Appliquées*, **104**(2):227–257, 2015. doi:10.1016/j.matpur.2015.02.006.

[CLP17] E. Croot, V.L. Lev, and P.P. Pach. Progression-free sets in $\mathbb{Z}_4^n$ are exponentially small. *Annals of Mathematics, Second series*, **185**(1):331–337, 2017. doi:10.4007/annals.2017.185.1.7.

[Con07] J.B. Conway. *A Course in Functional Analysis*, volume 96 of *Graduate Texts in Mathematics*. Springer, second edition, 2007.

[Day62] M.M. Day. *Normed Linear Spaces*. Ergebnisse der Mathematik und ihrer Grenzgebiete. Springer, 1962. Second printing, corrected.

[Der22] H. Derksen. The g-stable rank for tensors and the cap set problem. *Algebra & Number Theory*, **16**(5):1071–1097, 2022. doi:10.2140/ant.2022.16.1071.

[DF93] A. Defant and K. Floret. *Tensor Norms and Operator Ideals*, volume 176 of *Mathematics Studies*. North-Holland, 1993.

[DG20] J. van Dobben de Bruyn and D. Gijswijt. Treewidth is a lower bound on graph gonality. *Algebraic Combinatorics*, **3**(4):941–953, 2020. doi:10.5802/alco.124.

[DG21] J. van Dobben de Bruyn and D. Gijswijt. On the size of subsets of $\mathbb{F}_q^n$ avoiding solutions to linear systems with many linearly dependent columns, 2021. Preprint. URL: https://arxiv.org/abs/2111.09879.

[Dha90] D. Dhar. Self-organized critical state of sandpile automaton models. *Physical Review Letters*, **64**(14):1613–1616, 1990. doi:10.1103/PhysRevLett.64.1613.

[Dob12] J. van Dobben de Bruyn. Reduced divisors and gonality in finite graphs, 2012. Bachelor thesis, Leiden University.

[Dob20a] J. van Dobben de Bruyn. Representations and semisimplicity of ordered topological vector spaces, 2020. Preprint. URL: https://arxiv.org/abs/2009.11777.

[Dob20b] J. van Dobben de Bruyn. Tensor products of convex cones, 2020. Preprint. URL: https://arxiv.org/abs/2009.11843.

[Dob22] J. van Dobben de Bruyn. Almost all positive continuous linear functionals can be extended. *Positivity*, **26**(1):#15, 2022.

[DS70] Y. Dermenjian and J. Saint–Raymond. Produit tensoriel de deux cones convexes saillants. *Séminaire Choquet (initation à l'analyse)*, **9**(20):1–6, 1969/70. URL: http://www.numdam.org/item/SC_1969-1970__9_2_A10_0/.

[DS89] J.D. Deuschel and D.W. Stroock. *Large Deviations*. Academic Press, Boston, 1989.

[DSW21] J. van Dobben de Bruyn, H. Smit, and M. van der Wegen. Code and figures to accompany the paper "Discrete and metric divisorial gonality can be different", 2021. URL: https://github.com/gonality/discrete-vs-metric-dgon, doi:10.5281/zenodo.7675182.

[DSW22] J. van Dobben de Bruyn, H. Smit, and M. van der Wegen. Discrete and metric divisorial gonality can be different. *Journal of Combinatorial Theory, Series A*, **189**:#105619, 2022. doi:10.1016/j.jcta.2022.105619.

[DV21a] J. Draisma and A. Vargas. Catalan-many tropical morphisms to trees; Part I: Constructions. *Journal of Symbolic Computation*, **104**:580–629, 2021. doi:10.1016/j.jsc.2020.09.005.

[DV21b] J. Draisma and A. Vargas. On the gonality of metric graphs. *Notices of the American Mathematical Society*, **68**(5):687–695, 2021.

[DZ98] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer, New York, second edition, 1998.

[Ede04] Y. Edel. Extensions of generalized product caps. *Designs, Codes and Cryptography*, **31**(1):5–14, 2004. doi:10.1023/A:1027365901231.

[EEH+22] M. Echavarria, M. Everett, R. Huang, L. Jacoby, R. Morrison, and B. Weber. On the scramble number of graphs. *Discrete Applied Mathematics*, **310**:43–59, 2022. doi:10.1016/j.dam.2021.12.009.

[EG17] J.S. Ellenberg and D. Gijswijt. On large subsets of $\mathbb{F}_q^n$ with no three-term arithmetic progression. *Annals of Mathematics, Second series*, **185**(1):339–343, 2017. doi:10.4007/annals.2017.185.1.8.

[EL23] C. Elsholtz and G.F. Lipnik. Exponentially larger affine and projective caps. *Mathematika*, **69**(1):232–249, 2023. doi:10.1112/mtk.12173.

[EP20] C. Elsholtz and P.P. Pach. Caps and progression-free sets in $\mathbb{Z}_m^n$. *Designs, Codes and Cryptography*, **88**(10):2133–2170, 2020. doi:10.1007/s10623-020-00769-0.

[FGR87] P. Frankl, R.L. Graham, and V. Rödl. On subsets of abelian groups with no 3-term arithmetic progressions. *Journal of Combinatorial Theory, Series A*, **45**(1):157–161, 1987. doi:10.1016/0097-3165(87)90053-7.

[FKO82] H. Furstenberg, Y. Katznelson, and D. Ornstein. The ergodic theoretical proof of Szemerédi's theorem. *Bulletin of the American Mathematical Society, New series*, **7**(3):527–552, 1982. doi:10.1090/S0273-0979-1982-15052-2.

[FNT17] T. Fritz, T. Netzer, and A. Thom. Spectrahedral containment and operator systems with finite-dimensional realization. *SIAM Journal on Applied Algebra and Geometry*, **1**(1):556–574, 2017.

[Fre72] D.H. Fremlin. Tensor products of Archimedean vector lattices. *American Journal of Mathematics*, **94**(3):777–798, 1972.

[Fre74] D.H. Fremlin. Tensor products of Banach lattices. *Mathematische Annalen*, **211**(2):87–106, 1974.

[FT79] D.H. Fremlin and M. Talagrand. A decomposition theorem in tensor products of Archimedean vector lattices. *Mathematika*, **26**(2):302–305, 1979.

[Fur77] H. Furstenberg. Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. *Journal d'Analyse Mathématique*, **31**:204–256, 1977. doi:10.1007/BF02813304.

[Gij21] D. Gijswijt. Excluding affine configurations over a finite field, 2021. Preprint. URL: https://arxiv.org/abs/2112.12620.

[GK08] A. Gathmann and M. Kerber. A Riemann–Roch theorem in tropical geometry. *Mathematische Zeitschrift*, **259**(1):217–230, 2008. doi:10.1007/s00209-007-0222-4.

[GK10] O. van Gaans and A. Kalauch. Tensor products of Archimedean partially ordered vector spaces. *Positivity*, **14**(4):705–714, 2010.

[GL88] J.J. Grobler and C.C.A. Labuschagne. The tensor product of Archimedean ordered vector spaces. *Mathematical Proceedings of the Cambridge Philosophical Society*, **104**(2):331–345, 1988.

[GL89] J.J. Grobler and C.C.A. Labuschagne. An f-algebra approach to the Riesz tensor product of Archimedean Riesz spaces. *Quaestiones Mathematicae*, **12**(4):425–438, 1989. doi:10.1080/16073606.1989.9632194.

[GLMP04] Y. Gordon, A.E. Litvak, M. Meyer, and A. Pajor. John's decomposition in the general case and applications. *Journal of Differential Geometry*, **68**(1):99–119, 2004.

[Gow01] W.T. Gowers. A new proof of Szemerédi's theorem. *Geometric and Functional Analysis*, **11**(3):465–588, 2001. doi:10.1007/s00039-001-0332-9.

[Gow07] W.T. Gowers. Hypergraph regularity and the multidimensional Szemerédi theorem. *Annals of Mathematics, Second series*, **166**:897–946, 2007. doi:10.4007/annals.2007.166.897.

[Gow98] W.T. Gowers. A new proof of Szemerédi's theorem for arithmetic progressions of length four. *Geometric and Functional Analysis*, **8**(3):529–551, 1998. doi:10.1007/s000390050065.

[GPT01] A. Giannopoulos, I. Perissinaki, and A. Tsolomitis. John's theorem for an arbitrary pair of convex bodies. *Geometriae Dedicata*, **84**:63–79, 2001.

[Gro55] A. Grothendieck. Produits tensoriels topologiques et espaces nucléaires. *Memoirs of the American Mathematical Society*, **16**, 1955.

[GSW20] D. Gijswijt, H. Smit, and M. van der Wegen. Computing graph gonality is hard. *Discrete Applied Mathematics*, **287**:134–149, 2020. doi:10.1016/j.dam.2020.08.013.

[GT08] B. Green and T. Tao. The primes contain arbitrarily long arithmetic progressions. *Annals of Mathematics, Second series*, **167**(2):481–547, 2008. doi:10.4007/annals.2008.167.481.

[GW11] W.T. Gowers and J. Wolf. Linear forms and higher-degree uniformity for functions on $\mathbb{F}_p^n$. *Geometric and Functional Analysis*, **21**(1):36–69, 2011. doi:10.1007/s00039-010-0106-3.

[Hen18] K. Hendrey. Sparse graphs of high gonality. *SIAM Journal on Discrete Mathematics*, **32**(2):1400–1407, 2018. doi:10.1137/16M1095329.

[HFP76] E. Haynsworth, M. Fiedler, and V. Pták. Extreme operators on polyhedral cones. *Linear Algebra and its Applications*, **13**(1–2):163–172, 1976.

[Hil08] R. Hildebrand. Semidefinite descriptions of low-dimensional separable matrix cones. *Linear Algebra and its Applications*, **429**(4):901–932, 2008.

[HKN13] J. Hladký, D. Kráľ, and S. Norine. Rank of divisors on tropical curves. *Journal of Combinatorial Theory, Series A*, **120**(7):1521–1538, 2013. doi:10.1016/j.jcta.2013.05.002.

[HN21] B. Huber and T. Netzer. A note on non-commutative polytopes and polyhedra. *Advances in Geometry*, **21**(1):119–124, 2021. doi:10.1515/advgeom-2020-0029.

[Hol00] F. den Hollander. *Large Deviations*, volume 14 of *Fields Institute Monographs*. American Mathematical Society, Providence, RI, 2000.

[HP68] A. Hulanicki and R.R. Phelps. Some applications of tensor products of partially-ordered linear spaces. *Journal of Functional Analysis*, **2**(2):177–201, 1968.

[Jen21] D. Jensen. Chip firing and algebraic curves. *Notices of the American Mathematical Society*, **68**(11):1875–1881, 2021. doi:10.1090/noti2378.

[Jia21] Z. Jiang. Improved explicit upper bounds for the cap set problem, 2021. Preprint. URL: https://arxiv.org/abs/2103.06481.

[Joh48] F. John. Extremum problems with inequalities as subsidiary conditions. In *Studies and Essays presented to R. Courant on his 60th Birthday*, pages 187–204. Interscience, 1948.

[Kad51] R.V. Kadison. A representation theory for commutative topological algebra. *Memoirs of the American Mathematical Society*, **7**, 1951.

[Kal21] O. Kallenberg. *Foundations of Modern Probability*. Springer, Cham, third edition, 2021.

[Kal78] N.J. Kalton. Quotients of F-spaces. *Glasgow Mathematical Journal*, **19**(2):103–108, 1978.

[Kle08] A. Klenke. *Probability Theory*. Springer, London, 2008.

[Kle59] V. Klee. Some new results on smoothness and rotundity in normed linear space. *Mathematische Annalen*, **139**(1):51–63, 1959.

[KMZ20] S. Kopparty, G. Moshkovitz, and J. Zuiddam. Geometric rank of tensors and subrank of matrix multiplication. In S. Saraf, editor, *35th Computational Complexity Conference (CCC 2020)*, pages 35:1–35:21, Dagstuhl, Germany, 2020. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.CCC.2020.35.

[Köt79] G. Köthe. *Topological Vector Spaces II*, volume 237 of *Grundlehren der mathematischen Wissenschaften*. Springer, New York, NY, 1979.

[Köt83] G. Köthe. *Topological Vector Spaces I*, volume 159 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, second revised printing edition, 1983.

[KSS18] R. Kleinberg, D.E. Speyer, and W. Sawin. The growth rate of tri-colored sum-free sets. *Discrete Analysis*, **2018:12**:10pp., 2018. doi:10.19086/da.3734.

[Lam17] L. Lami. Non-classical correlations in quantum mechanics and beyond, 2017. Ph.D. thesis, Universitat Autònoma de Barcelona.

[Lan02] S. Lang. *Algebra*, volume 211 of *Graduate Texts in Mathematics*. Springer, New York, revised third edition, 2002.

[Lov19] S. Lovett. The analytic rank of tensors and its applications. *Discrete Analysis*, **2019:7**:10pp., 2019. doi:10.19086/da.8654.

[Luo11] Y. Luo. Rank-determining sets of metric graphs. *Journal of Combinatorial Theory, Series A*, **118**(6):1775–1793, 2011. doi:10.1016/j.jcta.2011.03.002.

[Man22] M. Manjunath. Brill–Noether existence on graphs via $\mathbb{R}$-divisors, polytopes and lattices. *Selecta Mathematica, New Series*, **28**(2):#35, 2022. doi:10.1007/s00029-021-00728-0.

[Mer64] H. Merklen. Producto tensorial de espacios vectoriales ordenados. *Notas de Matemáticas (Universidad Nacional de Ingeniería de Perú)*, **2**(2):41–57, 1964.

[Mes95] R. Meshulam. On subsets of finite abelian groups with no 3-term arithmetic progressions. *Journal of Combinatorial Theory, Series A*, **71**(1):168–172, 1995. doi:10.1016/0097-3165(95)90024-1.

[MG07] J. Matoušek and B. Gärtner. *Understanding and Using Linear Programming*. Springer, 2007. doi:10.1007/978-3-540-30717-4.

[MP14] B.D. McKay and A. Piperno. Partical graph isomorphism, II. *Journal of Symbolic Computation*, **60**:94–112, 2014. doi:10.1016/j.jsc.2013.09.003.

[MP20] B.D. McKay and A. Piperno. nauty and Traces, version 2.7r1, 2020. URL: https://pallini.di.uniroma1.it.

[MT19a] M. Mimura and N. Tokushige. Avoiding a star of three-term arthmetic progressions, 2019. Preprint. URL: https://arxiv.org/abs/1909.10507.

[MT19b] M. Mimura and N. Tokushige. Avoiding a shape, and the slice rank method for a system of equations, 2019. Preprint. URL: https://arxiv.org/abs/1909.10509.

[MT20] M. Mimura and N. Tokushige. Solving linear equations in a vector space over a finite field II, 2020. Preprint. URL: http://www.cc.u-ryukyu.ac.jp/~hide/sol2.pdf.

[MT22] R. Morrison and L. Tolley. Computing higher graph gonality is hard, 2022. Preprint. URL: https://arxiv.org/abs/2208.03573.

[Mül21] M.P. Müller. Probabilistic theories and reconstructions of quantum theory. *SciPost Physics Lecture Notes*, **28**:41pp., 2021. doi:10.21468/SciPostPhysLectNotes.28.

[Mul97] B. Mulansky. Tensor products of convex cones. In G. Nürnberger, J.W. Schmidt, and G. Walz, editors, *Multivariate Approximation and Splines*, volume 125 of *International Series on Numerical Mathematics*. Birkhäuser, Basel, 1997. doi:10.1007/978-3-0348-8871-4_14.

[MZ08] G. Mikhalkin and I. Zharkov. Tropical curves, their Jacobians and Theta functions. In V. Alexeev, A. Beauville, C.H. Clemens, and E. Izadi, editors, *Curves and Abelian Varieties*, volume 465 of *Contemporary Mathematics*, pages 203–230. American Mathematical Society, Providence, Rhode Island, 2008. doi:10.1090/conm/465.

[Nas20a] E. Naslund. Exponential bounds for the Erdős–Ginzburg–Ziv constant. *Journal of Combinatorial Theory, Series A*, **174**:105185, 2020. doi:10.1016/j.jcta.2019.105185.

[Nas20b] E. Naslund. The partition rank of a tensor and k-right corners in $\mathbb{F}_q^n$. *Journal of Combinatorial Theory, Series A*, **174**:#105190, 2020. doi:j.jcta.2019.105190.

[Nas22] E. Naslund. Upper bounds for families without weak Delta-systems, 2022. Preprint. URL: https://arxiv.org/abs/2203.13370.

[Nie82] N.J. Nielsen. The ideal property of tensor products of Banach lattices with applications to the local structure of spaces of absolutely summing operators. *Studia Mathematica*, **74**(3):247–272, 1982.

[Nie88] N.J. Nielsen. The positive approximation property of Banach lattices. *Israel Journal of Mathematics*, **62**(1):99–112, 1988.

[Nor19] S. Norin. A distribution on triples with maximum entropy marginal. *Forum of Mathematics, Sigma*, **7**:#E46, 2019. doi:10.1017/fms.2019.47.

[NRS06] B. Nagle, V. Rödl, and M. Schacht. The counting lemma for regular k-uniform hypergraphs. *Random Structures & Algorithms*, **28**(2):113–179, 2006. doi:10.1002/rsa.20117.

[NS17] E. Naslund and W. Sawin. Upper bounds for sunflower-free sets. *Forum of Mathematics, Sigma*, **5**:#E15, 2017. doi:10.1017/fms.2017.12.

[Peb18] L. Pebody. Proof of a conjecture of Kleinberg–Sawin–Speyer. *Discrete Analysis*, **2018:3**:7pp., 2018. doi:10.19086/da.3733.

[Per67] A.L. Peressini. *Ordered Topological Vector Spaces*. Harper's Series in Modern Mathematics. Harper & Row, 1967.

[Plá21] M. Plávala. General probabilistic theories: An introduction, 2021. Preprint. URL: https://arxiv.org/abs/2103.07469.

[Poo75] M.W. Poole. Structure properties of polyhedral cones, 1975. PhD thesis, Auburn University.

[Pop68] N. Popa. Produits tensoriels ordonnés. *Revue Roumaine de Mathématiques Pures et Appliquées*, **13**(2):235–246, 1968.

[Pop69] N. Popa. Sur les produits tensoriels ordonnés. *Mathematische Zeitschrift*, **110**(3):206–210, 1969.

[PS69] A.L. Peressini and D.R. Sherbert. Ordered topological tensor products. *Proceedings of the London Mathematical Society, 3rd series*, **19**(1):177–190, 1969.

[PSS18] B. Passer, O.M. Shalit, and B. Solel. Minimal and maximal matrix convex sets. *Journal of Functional Analysis*, **274**(11):3197–3253, 2018.

[PTT11] V.I. Paulsen, I.G. Todorov, and M. Tomforde. Operator system structures on ordered spaces. *Proceedings of the London Mathematical Society, 3rd series*, **102**(1):25–49, 2011.

[Roc70] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[Rot52] K. Roth. Sur quelques ensembles d'entiers. *Comptes rendus hebdomadaires des séances de l'Académie des Sciences*, **234**:388–390, 1952.

[Rot53] K.F. Roth. On certain sets of integers. *Journal of the London Mathematical Society*, **s1-28**(1):104–109, 1953. doi:10.1112/jlms/s1-28.1.104.

[RS04] V. Rödl and J. Skokan. Regularity lemma for k-uniform hypergraphs. *Random Structures & Algorithms*, **25**(1):1–42, 2004. doi:10.1002/rsa.20017.

[RS06] V. Rödl and J. Skokan. Applications of the regularity lemma for uniform hypergraphs. *Random Structures & Algorithms*, **28**(2):180–194, 2006. doi:10.1002/rsa.20108.

[RS15] F. Rassoul-Agha and T. Seppäläinen. *A Course on Large Deviations with an Introduction to Gibbs Measures*, volume 162 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2015.

[RS82] W.M. Ruess and C.P. Stegall. Extreme points in duals of operator spaces. *Mathematische Annalen*, **261**(4):535–546, 1982.

[RS86a] N. Robertson and P.D. Seymour. Graph minors. II. Algorithmic aspects of tree-width. *Journal of Algorithms*, **7**(3):309–322, 1986. doi:10.1016/0196-6774(86)90023-4.

[RS86b] W.M. Ruess and C.P. Stegall. Weak*-denting points in duals of operator spaces. In N.J. Kalton and E. Saab, editors, *Banach Spaces, Proceedings of the Missouri Conference held in Columbia, USA, June 24–29, 1984*, volume 1166 of *Lecture Notes in Mathematics*, pages 158–169. Springer, 1986.

[Rud91] W. Rudin. *Functional Analysis*. McGraw–Hill, second edition, 1991.

[Ruz93] I.Z. Ruzsa. Solving a linear equation in a set of integers I. *Acta Arithmetica*, **65**(3):259–282, 1993. doi:10.4064/aa-65-3-259-282.

[Ruz95] I.Z. Ruzsa. Solving a linear equation in a set of integers II. *Acta Arithmetica*, **72**(4):385–397, 1995. doi:10.4064/aa-72-4-385-397.

[Rya02] R.A. Ryan. *Introduction to Tensor Products of Banach Spaces*. Springer Monographs in Mathematics. Springer, London, 2002.

[Sau21] L. Sauermann. On the size of subsets of $\mathbb{F}_p^n$ without p distinct elements summing to zero. *Israel Journal of Mathematics*, **243**(1):63–79, 2021. doi:10.1007/s11856-021-2145-x.

[Sau22] L. Sauermann. Finding solutions with distinct variables to systems of linear equations over $\mathbb{F}_p^n$. *Mathematische Annalen*, 2022. Online first articles. doi:10.1007/s00208-022-02391-y.

[Saw18] W. Sawin. Bounds for matchings in nonabelian groups. *Electronic Journal of Combinatorics*, **25**(4):P4.23, 2018. doi:10.37236/7520.

[Sch58] H. Schaefer. Halbgeordnete lokalkonvexe Vektorräume. *Mathematische Annalen*, **135**(2):115–141, 1958.

[Sch60] H. Schaefer. Halbgeordnete lokalkonvexe Vektorräume III. *Mathematische Annalen*, **141**(2):113–142, 1960.

[Sch72] H.H. Schaefer. Normed tensor products of Banach lattices. *Israel Journal of Mathematics*, **13**(4):400–415, 1972.

[Sch74] H.H. Schaefer. *Banach Lattices and Positive Operators*. Grundlehren der mathematischen Wissenschaften. Springer, 1974.

[Sch99] H.H. Schaefer. *Topological Vector Spaces*, volume 3 of *Graduate Texts in Mathematics*. Springer, second edition, 1999.

[Sha21] O.M. Shalit. Dilation theory: A guided tour. In M.A. Bastos, L. Castro, and A.Y. Karlovich, editors, *Operator Theory, Functional Analysis and Applications*, pages 551–623. Birkhäuser, Cham, 2021.

[ST90] C.H. Sung and B.S. Tam. A study of projectionally exposed cones. *Linear Algebra and its Applications*, **139**:225–252, 1990.

[ST93] P.D. Seymour and R. Thomas. Graph searching and a min–max theorem for tree-width. *Journal of Combinatorial Theory, Series B*, **58**(1):22–33, 1993. doi:10.1006/jctb.1993.1027.

[SW70] J. Stoer and C. Witzgall. *Convexity and Optimization in Finite Dimensions I*, volume 163 of *Die Grundlehren der mathematischen Wissenschaften*. Springer, 1970.

[Sze69] E. Szemerédi. On sets of integers containing no four elements in arithmetic progression. *Acta Mathematica Academiae Scientiarum Hungaricae*, **20**(1-2):89–104, 1969. doi:10.1007/BF01894569.

[Sze75] E. Szemerédi. On sets of integers containing no k elements in arithmetic progression. *Acta Arithmetica*, **27**:199–245, 1975. doi:10.4064/aa-27-1-199-245.

[Tam77a] B.S. Tam. Some aspects of finite dimensional cones, 1977. PhD thesis, University of Hong Kong.

[Tam77b] B.S. Tam. Some results of polyhedral cones and simplicial cones. *Linear and Multilinear Algebra*, **4**(4):281–284, 1977.

[Tam85] B.S. Tam. On the duality operator of a convex cone. *Linear Algebra and its Applications*, **64**:33–56, 1985.

[Tam92] B.S. Tam. On the structure of the cone of positive operators. *Linear Algebra and its Applications*, **167**:65–85, 1992.

[Tam95] B.S. Tam. Extreme positive operators on convex cones. In K.Y. Chan and M.C. Liu, editors, *Five Decades as a Mathematician and Educator: On the 80th Birthday of Professor Yung-Chow Wong*, pages 515–558. World Scientific, Singapore, 1995.

[Tao10] T. Tao. Answer to the MathOverlflow question 'How to recognise that the polynomial method might work', 2010. URL: `https://mathoverflow.net/a/43549`.

[Tao16] T. Tao. A symmetric formulation of the Croot–Lev–Pach–Ellenberg–Gijswijt capset bound, 2016. Blog post. URL: `https://terrytao.wordpress.com/2016/05/18/a-symmetric-formulation-of-the-croot-lev-pach-ellenberg-gijswijt-capset-bound/`.

[TS16] T. Tao and W. Sawin. Notes on the "slice rank" of tensors, 2016. Blog post. URL: `https://terrytao.wordpress.com/2016/08/24/notes-on-the-slice-rank-of-tensors/`.

[Tse76] I.I. Tseitlin. The extreme points of the unit ball of certain spaces of operators. *Matematicheskie Zametki*, **20**(4):521–527, 1976.

[TV06] T. Tao and V.H. Vu. *Additive Combinatorics*, volume 105 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 2006.

[Tyr22] F. Tyrrell. New lower bounds for cap sets, 2022. Preprint. URL: `http://arxiv.org/abs/2209.10045v1`.

[Ura00] H. Urakawa. A discrete analogue of the harmonic morphism and Green kernel comparison theorems. *Glasgow Mathematical Journal*, **42**(3):319–334, 2000. `doi:10.1017/S0017089500030019`.

[Var84] S.R.S. Varadhan. *Large Deviations and Applications*. SIAM, Philadelphia, 1984.

[Wae27] B.L. van der Waerden. Beweis einer Baudetschen Vermutung. *Nieuw Archief voor Wiskunde, Tweede reeks*, **15**:212–216, 1927.

[Wei12] S. Weis. Duality of non-exposed faces. *Journal of Convex Analysis*, **19**(3):815–835, 2012.

[Wer87] D. Werner. Denting points in tensor products of Banach spaces. *Proceedings of the American Mathematical Society*, **101**(1):122–126, 1987.

[Wit74] G. Wittstock. Eine Bemerkung über Tensorprodukte von Banachverbänden. *Archiv der Mathematik*, **25**(1):627–634, 1974.

[Wor19] M. Wortel. Lexicographic cones and the ordered projective tensor product. In G. Buskes, M. de Jeu, P. Dodds, A. Schep, F. Sukochev, J. van Neerven, and A. Wickstead, editors, *Positivity and Noncommutative Analysis: Festschrift in Honour of Ben de Pagter on the Occasion of his 65th Birthday*, pages 601–609. Birkhäuser, Cham, 2019. `doi:10.1007/978-3-030-10850-2_30`.

[Zaa97] A.C. Zaanen. *Introduction to Operator Theory in Riesz Spaces*. Springer, 1997.

[Zam87] T. Zamfirescu. Nearly all convex bodies are smooth and strictly convex. *Monatshefte für Mathematik*, **103**(1):57–62, 1987.

[Zui18] J. Zuiddam. Algebraic complexity, asymptotic spectra and entanglement polytopes, 2018. PhD thesis, University of Amsterdam.

Summary

This dissertation addresses three different topics on the interface of combinatorics, algebra, and geometry.

In Part I, we study *divisorial gonality* of graphs, a relatively new graph parameter which has its roots in algebraic geometry. This parameter dates back to around 2007, when Baker initiated a programme to translate results from classical algebraic geometry into statements about chip-firing games on finite graphs. By then, it had been known for a while that graphs behave in many ways as discrete analogues of Riemann surfaces (e.g. [BHN97, Ura00]), but Baker took this one step further by providing a concrete way to specialize divisors from curves to graphs [Bak08], and by formulating and proving (together with Norine) a combinatorial analogue of the classical Riemann–Roch theorem from geometry [BN07]. This ushered in a period of fruitful interplay between algebraic geometry, tropical geometry, and graph theory, whose highlights include various combinatorial Riemann–Roch theorems [BN07, MZ08, GK08, AM10, AC13, CLM15, Bac17], a combinatorial proof of the Brill–Noether non-existence theorem in algebraic geometry [CDPR12], and unexpected connections with structural graph theory and parametrized complexity [DG20, BCW22b].

In this dissertation, we make two contributions to the theory of divisors on graphs. First, besides its connections with algebraic geometry, divisorial gonality of graphs is also closely related to *treewidth*, a graph parameter that plays an important role in structural graph theory and parametrized complexity. In 2014, Gijswijt and the author showed that treewidth is a lower bound for the divisorial gonality [DG20], but the proof was not constructive. In this dissertation, we give a constructive proof of the same fact, by exhibiting a polynomial time algorithm that converts a positive rank effective divisor of degree k into a tree decomposition of width at most k. This sheds new light on the connection between gonality and treewidth, and makes it easy to apply dynamic programming techniques from parametrized complexity in the context of graphs with bounded gonality.

Second, we look at the Brill–Noether conjecture for graphs, originally formulated by Baker [Bak08] as a combinatorial analogue of classical Brill–Noether theory. This conjecture consists of two parts: an 'existence' and a 'non-existence' statement. The non-existence part was settled by Cools, Draisma, Payne and Robeva [CDPR12], but the existence part is still wide open. Since Brill–Noether existence is known to be true for metric graphs, the most obvious approach towards a proof for graphs is to show

that the gonality of a graph is equal to the gonality of the associated metric graph with unit lengths. This was another conjecture of Baker [Bak08], called the *subdivision conjecture*. In this dissertation, we disprove the subdivision conjecture, by giving a family of examples where the gonality of the graph is strictly larger than the gonality of the associated metric graph. This shuts down the most obvious approach towards a proof of the Brill–Noether conjecture, and makes it unclear whether or not the latter is likely to be true. We have not been able to prove or disprove the Brill–Noether conjecture.

In Part II, we study an application of the *slice rank polynomial method* to the problem of avoiding affine configurations in subsets of $\mathbb{F}_q^n$. For a long time, it had been an open problem to determine whether or not there is a constant $c < 3$ such that every subset of $\mathbb{F}_3^n$ without non-trivial 3-term arithmetic progressions has size at most c^n. This problem, known as the *cap set problem*, was solved in 2016 by Ellenberg and Gijswijt [EG17], building on a new technique within the polynomial method introduced earlier that same year by Croot, Lev, and Pach [CLP17]. Their proof was subsequently recast by Tao [Tao16] in terms of a new rank function for tensors, called *slice rank*, and this new set of techniques is now known as the *slice rank polynomial method*. This method has been very successful in solving the cap set problem and a few related problems, but the more general problem of avoiding non-trivial k-term arithmetic progressions in subsets of $\mathbb{F}_p^n$ is still wide open for $k \geq 4$.

In this dissertation, we make partial progress towards this problem, by studying the broader problem of avoiding affine configurations in subsets of $\mathbb{F}_q^n$. Here, instead of avoiding non-trivial k-term arithmetic progressions, we seek to avoid non-trivial solutions to a system of balanced linear equations,

$$\begin{cases} a_{11}\boldsymbol{x_1} + \cdots + a_{1k}\boldsymbol{x_k} = 0, \\ \qquad\qquad \vdots \\ a_{m1}\boldsymbol{x_1} + \cdots + a_{mk}\boldsymbol{x_k} = 0; \end{cases} \tag{$\star$}$$

where $a_{ij} \in \mathbb{F}_q$ are scalars such that $a_{i1} + \cdots + a_{ik} = 0$ for all $i \in [m]$, and $\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}$ are vectors in the affine space $\mathbb{F}_q^n$ as $n \to \infty$. If the number of variables is sufficiently large ($k \geq 2m + 1$), then a routine application of the slice rank method shows that there is a constant $C_{q,m,k} < q$ such that every subset $A \subseteq \mathbb{F}_q^n$ of size $|A| \geq (C_{q,m,k})^n$ contains a solution $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}) \in A^k$ of ($\star$) where the $\boldsymbol{x_i}$ are not all equal. Our contribution is that, for certain classes of balanced linear systems, we extend this to find a solution $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}) \in A^k$ of ($\star$) where the $\boldsymbol{x_i}$ are pairwise distinct, or even maximally affinely independent (in the sense that $\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}$ do not satisfy any balanced linear equation over $\mathbb{F}_q$ that is not a linear combination of the rows of ($\star$)). This generalizes earlier results by Mimura and Tokushige [MT19a, MT19b, MT20], but was later superseded by a more general result by Gijswijt [Gij21].

In Part III, we study *tensor products of convex cones*. This topic has recently come up in many different areas of mathematics (and beyond), ranging from functional analysis and operator theory to approximation theory and theoretical physics. However, most of the existing literature focuses on one of two particular cases, namely

Archimedean lattice cones (in the functional analysis literature) or closed, proper and generating cones in finite-dimensional spaces (in linear algebra and most applications to other fields). This excludes most cones from being considered, including even standard cones such as infinite-dimensional positive semidefinite cones and lexicographical cones. For general cones, results are few and far between, and many basic questions remain unanswered.

In this dissertation, we address this gap in the literature by studying the problem in full generality. We generalize a few known results to the general case, and we prove many results which are altogether new. Our main contributions are the following: (i) We show that the projective/injective cone has mapping properties analogous to those of the projective/injective norm; (ii) We establish direct formulas for the lineality space of the projective/injective cone, in particular providing necessary and sufficient conditions for the cone to be proper; (iii) We prove that the projective/injective tensor product of two closed proper cones is contained in a closed proper cone; (iv) We show how to construct faces of the projective/injective cone from faces of the base cones, in particular providing a complete characterization of the extremal rays of the projective cone. As an application, we also show that the tensor product of two symmetric convex sets preserves proper faces; (v) For closed cones in finite-dimensional spaces, we show that the projective cone is closed, and almost always strictly contained in the injective cone, thereby confirming a conjecture of Barker for nearly all convex cones. As this dissertation was being written, this last result was superseded by simultaneous discovery by Aubrun, Lami, Palazuelos and Plávala [ALPP21], who independently proved Barker's conjecture in full generality (for all closed, proper and generating cones in finite-dimensional spaces). We recover their result for a large class of cones, using completely different techniques.

Samenvatting

Dit proefschrift gaat over drie verschillende onderwerpen op het raakvlak tussen combinatoriek, algebra en meetkunde.

In Deel I bestuderen we de *divisoriale gonaliteit* van grafen, een relatief jonge graafparameter die zijn wortels heeft in de algebraïsche meetkunde. Deze parameter vond zijn oorsprong omstreeks 2007, toen Baker een onderzoek opzette dat probeert resultaten uit de klassieke algebraïsche meetkunde te vertalen naar uitspraken over een spel met fiches op een eindige graaf ('chip-firing games'). Het was destijds al enige tijd bekend dat grafen zich op meerdere manieren gedragen als het discrete analogon van Riemannoppervlakken (bijv. [BHN97, Ura00]), maar Baker ging nog een stap verder door een concrete manier te geven om divisoren te 'specialiseren' van krommen naar grafen [Bak08], en door (samen met Norine) een combinatorisch analogon van de klassieke Riemann–Roch stelling uit de meetkunde te formuleren en bewijzen [BN07]. Dit leidde een periode in van succesvolle kruisbestuiving tussen algebraïsche meetkunde, tropische meetkunde en grafentheorie, met als hoogtepunten onder meer verschillende combinatorische Riemann–Roch stellingen [BN07, MZ08, GK08, AM10, AC13, CLM15, Bac17], een combinatorisch bewijs van de Brill–Noether non-existentie stelling uit de algebraïsche meetkunde [CDPR12], en onverwachte dwarsverbanden met structurele grafentheorie en geparametriseerde complexiteit [DG20, BCW22b].

In dit proefschrift leveren we twee bijdragen aan de theorie van divisoren op grafen. Ten eerste: naast zijn verband met algebraïsche meetkunde is divisoriale gonaliteit ook nauw verwant aan *boombreedte*, een graafparameter die een belangrijke rol speelt in de structurele grafentheorie en de geparametriseerde complexiteit. In 2014 bewezen Gijswijt en de auteur van dit proefschrift dat de boombreedte een ondergrens is voor de divisoriale gonaliteit [DG20], maar het bewijs was niet-constructief. In dit proefschrift geven we een constructief bewijs van hetzelfde resultaat. Dat doen we door een algoritme te geven dat een gegeven effectieve divisor van positieve rang en graad k in polynomiale tijd omzet naar een boomdecompositie van breedte hoogstens k. Dit geeft een nieuw inzicht in de relatie tussen gonaliteit en boombreedte, en maakt het bovendien eenvoudig om bestaande algoritmen uit de geparametriseerde complexiteit, gebaseerd op dynamisch programmeren op een boomdecompositie, toe te passen op grafen met begrensde gonaliteit.

Ten tweede bestuderen we het Brill–Noether vermoeden voor grafen, oorspronkelijk geformuleerd door Baker [Bak08] als een combinatorisch analogon van klassieke Brill–

Noether theorie. Dit vermoeden bestaat uit twee delen: een 'existentie' en een 'non-existentie' gedeelte. Het non-existentie gedeelte is opgelost door Cools, Draisma, Payne en Robeva [CDPR12], maar het existentie gedeelte is nog steeds open. Gezien Brill–Noether existentie bewezen is voor metrische grafen, is de meest voor de hand liggende strategie om te bewijzen dat de gonalilteit van een graaf gelijk is aan de gonaliteit van de bijbehorende metrische graaf waarin iedere zijde lengte 1 heeft. Dit is eveneens een vermoeden van Baker [Bak08], dat we het *onderverdelingsvermoeden* zullen noemen. In dit proefschrift geven we een tegenvoorbeeld voor het onderverdelingsvermoeden door een familie van voorbeelden te construeren waarin de gonaliteit van de graaf strikt groter is dan de gonaliteit van de bijbehorende metrische graaf. Dit sluit de meest voor de hand liggende route naar een bewijs van het Brill–Noether vermoeden af, en maakt het bovendien onduidelijk of het laatstgenoemde überhaupt waar is. We zijn er niet in geslaagd om een bewijs of tegenvoorbeeld voor het Brill–Noether vermoeden te vinden.

In Deel II bestuderen we een toepassing van de *slice rank methode* op het begrenzen van de maximale grootte van een verzameling van $\mathbb{F}_q^n$ waarin bepaalde affiene configuraties worden vermeden. Gedurende lange tijd was het een open probleem om te bepalen of er een constante $c < 3$ bestaat zodat elke deelverzameling van $\mathbb{F}_3^n$ van grootte minstens c^n een niet-triviale rekenkundige rij van lengte 3 bevat. Dit probleem, dat bekend staat als het *cap set probleem*, werd in 2016 opgelost door Ellenberg en Gijswijt [EG17], door voort te borduren op een nieuwe techniek binnen de polynomiale methode die eerder dat jaar was geïntroduceerd door Croot, Lev en Pach [CLP17]. Hun bewijs werd vervolgens herschreven door Tao [Tao16] in termen van een nieuwe rangfunctie voor tensoren, genaamd *slice rank* ('plakjesrang'), en sindsdien staan deze technieken bekend als de *slice rank methode*. Deze methode is zeer succesvol gebleken in het oplossen van het cap set probleem en een aantal gerelateerde problemen, maar het bredere probleem van vermijden van niet-triviale rekenkundige rijen van lengte k in deelverzamelingen van $\mathbb{F}_p^n$ is nog steeds open voor $k \geq 4$.

In dit proefschrift maken enige vooruitgang bij dit probleem door te kijken naar het algemenere probleem van vermijden van affiene configuraties in deelverzamelingen van $\mathbb{F}_q^n$. In plaats van een rekenkundige rij van lengte k willen we nu niet-triviale oplossingen van een stelsel van gebalanceerde lineaire vergelijkingen vermijden. Met andere woorden, we hebben een stelsel

$$\begin{cases} a_{11}\boldsymbol{x_1} + \cdots + a_{1k}\boldsymbol{x_k} = 0, \\ \qquad\qquad \vdots \\ a_{m1}\boldsymbol{x_1} + \cdots + a_{mk}\boldsymbol{x_k} = 0; \end{cases} \tag{$\star$}$$

met $a_{ij} \in \mathbb{F}_q$ scalairen zodat $a_{i1} + \cdots + a_{ik} = 0$ voor alle $i \in [m]$, en met $\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}$ vectoren in de affiene ruimte $\mathbb{F}_q^n$, waarbij $n \to \infty$. Als het aantal variabelen voldoende groot is ($k \geq 2m + 1$), dan kan men middels een eenvoudige toepassing van de slice rank methode laten zien dat er een constante $C_{q,m,k} < q$ bestaat zodat iedere deelverzameling $A \subseteq \mathbb{F}_q^n$ van grootte $|A| \geq (C_{q,m,k})^n$ een oplossing $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}) \in A^k$ van ($\star$) bevat waarin de vectoren $\boldsymbol{x_i}$ niet allemaal hetzelfde zijn. Onze bijdrage is dat we

dit voor bepaalde klassen van gebalanceerde lineaire stelsels uitbreiden om een oplossing $(\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}) \in A^k$ van $(\star)$ te vinden waarin de vectoren $\boldsymbol{x_i}$ paarsgewijs verschillend zijn, of zelfs maximaal affien onafhankelijk (in die zin dat de vectoren $\boldsymbol{x_1}, \ldots, \boldsymbol{x_k}$ aan geen enkele gebalanceerde lineaire vergelijking voldoen die niet een lineaire combinatie van de rijen van $(\star)$ is). Dit is een generalisatie van eerdere resultaten van Mimura en Tokushige [MT19a, MT19b, MT20], maar is sindsdien alweer verder veralgemeniseerd door Gijswijt [Gij21].

In Deel III bestuderen we *tensorproducten van convexe kegels*. Dit onderwerp is de afgelopen jaren langsgekomen in allerlei verschillende takken van wiskunde (en daarbuiten), variërend van functionaalanalyse en operatorentheorie tot benaderingstheorie en theoretische fysica. Desalniettemin richt het overgrote deel van de bestaande literatuur zich enkel op één van de volgende twee speciale gevallen: Archimedische traliekegels (in de functionaalanalylse), of gesloten, echte, voortbrengende kegels in eindig-dimensionale ruimten (in de lineaire algebra en in de meeste toepassingen in andere vakgebieden). Hierdoor worden de meeste kegels buiten beschouwing gelaten, waaronder zelfs standaardkegels zoals oneindig-dimensionale positief semidefiniete kegels en lexicografische kegels. Voor algemene kegels zijn de resultaten schaars en zijn veel basisvragen vooralsnog onbeantwoord.

In dit proefschrift vullen we dit gat in de literatuur door het probleem zo algemeen mogelijk te bestuderen. We generaliseren een aantal bekende resultaten naar het algemene geval en we bewijzen veel resultaten die überhaupt nieuw zijn. Onze belangrijkste bijdragen zijn als volgt: (i) We bewijzen dat de projectieve/injectieve kegel afbeeldingseigenschappen heeft die analoog zijn aan die van de projectieve/injectieve norm; (ii) We geven directe formules voor de ruimte van linealiteit van de projectieve/injectieve kegel, en geven daarmee in het bijzonder noodzakelijke en voldoende voorwaarden voor het echt zijn van de kegel; (iii) We bewijzen dat het projectieve/injectieve tensorproduct van twee gesloten echte kegels bevat is in een gesloten echte kegel; (iv) We laten zien hoe men zijvlakken van de projectieve/injectieve kegel kan construeren uit zijvlakken van de oorspronkelijke kegels, wat in het bijzonder leidt tot een volledige beschrijving van de extremale halflijnen van de projectieve kegel. Als toepassing hiervan laten we tevens zien dat het tensorproduct van symmetrische convexe verzamelingen niet-triviale zijvlakken bewaart; (v) Voor gesloten kegels in eindig-dimensionale vectorruimten bewijzen we dat de projectieve kegel gesloten is, en nagenoeg altijd strikt bevat in de injectieve kegel. Hiermee bewijzen we een vermoeden van Barker voor nagenoeg alle convexe kegels. Terwijl dit proefschrift werd geschreven, werd dit laatste resultaat overtroffen door gelijktijdig werk van Aubrun, Lami, Palazuelos en Plávala [ALPP21], die onafhankelijk van ons erin slaagden om Barker's vermoeden volledig te bewijzen (voor alle gesloten, echte, voortbrengende kegels in eindig-dimensionale vectorruimten). We bevestigen hun resultaat voor een grote klasse van convexe kegels, met totaal andere bewijzen.

Lijst van publicaties

Preprints

9. **Divisorial and geometric gonality of higher-rank tropical curves** (met David Holmes en David van der Vorm), 2022. arXiv:2112.04205v2.

8. **The minimal Archimedean order unitization of seminormed ordered vector spaces**, 2022. arXiv:2204.13688.

7. **On the size of subsets of $\mathbb{F}_q^n$ avoiding solutions to linear systems with repeated columns** (met Dion Gijswijt), 2021. arXiv:2111.09879.

6. **Tensor products of convex cones**, 2020. arXiv:2009.11843.

5. **Representations and semisimplicity of ordered topological vector spaces**, 2020. arXiv:2009.11777.

Gepubliceerde artikelen

4. **Constructing tree decompositions of graphs with bounded gonality** (met Hans Bodlaender, Dion Gijswijt en Harry Smit), Journal of Combinatorial Optimization, **44**(4):2681–2699, 2022. doi:10.1007/s10878-021-00762-w. Een eerdere versie was gepubliceerd als conference paper op COCOON 2020, doi:10.1007/978-3-030-58150-3\textunderscore31.

3. **Discrete and metric divisorial gonality can be different** (met Harry Smit en Marieke van der Wegen), Journal of Combinatorial Theory, Series A, **189**:#105619, 2022. doi:10.1016/j.jcta.2022.105619

2. **Almost all positive continuous linear functionals can be extended**, Positivity, **26**(1):#15, 2022. doi:10.1007/s11117-022-00881-6

1. **Treewidth is a lower bound on graph gonality** (met Dion Gijswijt), Algebraic Combinatorics **3**(4):941–953, 2020. doi:10.5802/alco.124

Proefschriften en scripties

- **Divisorial gonality of graphs, the slice rank polynomial method, and tensor products of convex cones**, Proefschrift, Technische Universiteit Delft, 2023.
- **Connections between the general theories of ordered vector spaces and C^*-algebras**, Masterscriptie, Universiteit Leiden, 2018.
- **Reduced divisors and gonality in finite graphs**, Bachelorscriptie, Universiteit Leiden, 2012.

Wetenschappelijke software

- **dgon-tools** (met Harry Smit en Marieke van der Wegen), software voor het berekenen van de divisoriale gonaliteit van grafen, 2021. `https://github.com/gonality/dgon-tools/`, `doi:10.5281/zenodo.7675184`.