

**From the holdings of the Department of Special Collections , Princeton University Library,
Princeton, NJ**

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff. For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website.

<https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines>. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. <https://library.princeton.edu/special-collections/ask-us>

Doctoral Dissertations, 2011-2026

Mathematics

2013

An assortment of results in combinatorics and compressed sensing

Alexeev, Boris

Princeton, NJ : Princeton University

<http://arks.princeton.edu/ark:/88435/dsp01nc580m775>

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. <https://www.copyright.gov/title17/title17.pdf>

AN ASSORTMENT OF RESULTS IN
COMBINATORICS AND COMPRESSED SENSING

BORIS ALEXEEV

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF
MATHEMATICS
ADVISER: JOHN HORTON CONWAY

SEPTEMBER 2013

© Copyright by Boris Alexeev, 2013.

All Rights Reserved

Abstract

We prove a variety of results spanning combinatorics, compressed sensing, and theoretical computer science. Our methods throughout are primarily combinatorial, but we also use algebraic techniques and occasionally estimates from analysis.

We begin by proving lower and upper bounds for tensor rank, a purely algebraic quantity of importance to complexity theory. As the highlight, we improve the best known lower bounds for explicit tensors.

We continue by proving some complexity-theoretic results of interest to compressed sensing. Specifically, given standard complexity-theoretic assumptions, we show that any natural extension of the Mumford-Shah functional is infeasible to compute, as is the problem of testing a frame for possessing full spark. Along the way, we both construct some explicit full spark frames and show that full spark frames are dense in the set of Parseval frames.

We follow the theme of merging complexity and compressed sensing further by using insights from the theory of expander graphs to give an efficient procedure to reconstruct a signal from intensity measurements, a problem known as phase retrieval. Our procedure improves upon the efficiency of all previously-known methods.

In the penultimate chapter, we complete the characterization of the basic classes in the proof of the strong perfect graph theorem.

Finally, we resolve a 75-year old conjecture in combinatorial number theory due to Richard Rado.

Acknowledgements

Given the context, I primarily wish to thank everyone who has contributed to my mathematical development, from my parents to all of my friends, teachers, and professors in school, college, graduate school, and in between. I've learned so much from and been shaped by so many mentors. I particularly owe a lot to Edward Davis, Shareef Bacchus, Steve Sigur, and Ted Shifrin. Of course, I have also been inspired by the broader community of mathematicians both through personal interactions and published articles.

I would like to thank my coauthors throughout the years that worked with me on all sorts of mathematics: Afonso Bandeira, Jameson Cahill, Matt Fickus, Michael Forbes, Jacob Fox, Sasha Fradkin, Ron Graham, Brian Jacokes, Ilhee Kim, Dustin Mixon, Jacob Tsimmerman, and Rachel Ward.

On a very personal level, thank you, Mara, for everything.

I have been blessed with wonderful friends throughout the past few years. Thank you especially, Shaye, Percy, Jacob, and Hannah.

Finally, I would like to thank the people of the Princeton math department, most significantly my advisor John Conway, for everything they've done for me over the past five years. Like many of my fellow graduate students, I would like to thank Jill LeClair for being tremendously helpful and everyone who frequented the common room for making it such a fun place.

To my parents, Valery and Natalia Alexeev

Contents

Abstract	iii
Acknowledgements	iv
1 Introduction	1
1.1 Background and overview of results	1
1.2 Tensor rank	1
1.3 The complexity of Mumford-Shaw type regularization	2
1.4 Full spark frames	3
1.5 Phase retrieval with polarization	3
1.6 Forbidden induced subgraphs of double-split graphs	4
1.7 Equations resolving a conjecture of Rado on partition regularity . . .	4
1.8 Publications	5
1.9 Original authorship	6
1.10 Further directions	6
2 Tensor rank: some lower and upper bounds	8
2.1 Introduction	8
2.2 Prior work	10
2.3 Our results	12
2.4 Definitions and notation	14
2.5 Combinatorially-defined tensors	15

2.6	Permutation tensors	21
2.6.1	Rank lower bounds	22
2.6.2	Rank upper bounds	22
2.7	Monotone tensor rank	32
2.8	Basic facts about tensors	34
2.9	Layer reduction	41
2.10	Algebraically-defined tensors	46
2.11	Higher-order tensors	51
2.12	Tightness of results	53
3	On the complexity of Mumford-Shah type regularization	56
3.1	Introduction	56
3.2	Truncated quadratic minimization reformulated	58
3.3	Reduction to SUBSET-SUM	60
3.4	Connection to sparse recovery	62
4	Full spark frames	64
4.1	Introduction	64
4.2	Deterministic constructions of full spark frames	69
4.3	Full spark frames are dense	87
4.4	The computational complexity of verifying full spark	92
5	Phase retrieval with polarization	100
5.1	Introduction	100
5.2	The noiseless case	102
5.3	The noisy case	110
5.4	Concluding remarks	119
5.5	Appendix	120
5.5.1	Graph pruning	120

5.5.2	Angular synchronization	121
5.5.3	Projective uniformity	127
5.5.4	Removing large vertices	133
5.5.5	Main result	135
6	Forbidden induced subgraphs of double-split graphs	140
6.1	Introduction	140
6.2	Almost-split graphs	143
6.3	Excluding 6 graphs	146
7	Equations resolving a conjecture of Rado on partition regularity	165
8	Notes	168

Chapter 1

Introduction

1.1 Background and overview of results

The results of this thesis span several different fields: computational complexity theory, compressed sensing, the theory of perfect graphs, and combinatorial number theory.

1.2 Tensor rank

Computational complexity theory, broadly speaking, studies how difficult it is to perform different tasks under different computational models. One particularly important model is that of *circuits*, which closely resemble the actual electric circuits underlying computer chips. As it happens, it has historically been challenging to prove strong *circuit lower bounds*. In this work, we do not directly work on this problem, nor even define the terms involved. Instead, in Chapter 2 we follow an approach suggested by the results of Strassen [Str73] and Raz [Raz10], which showed that good enough *tensor rank lower bounds* have implications for algebraic circuit (and formula) lower bounds.

Tensor rank is a generalization of matrix rank for higher-order “matrices”, and as

such is a purely algebraic quantity. However, the quantity is particularly complicated, and can reasonably be considered a computational model in its own right.

We explore tensor rank lower and upper bounds, focusing on explicit tensors. We construct explicit tensors whose rank matches the best known bounds in a few cases, and improves them in all other cases.

Inspired by the observation that permutation matrices are high rank, we explore generalized versions which we denote permutations tensors. We show, by counting, that there exist permutation tensors with high rank. However, we show that a natural class of permutation tensors (which we call group tensors) does not achieve these bounds. This eliminates perhaps the most natural candidates for high tensor rank.

We also explore *monotone tensor rank*, which is defined by natural analogy to other uses of *monotone* in complexity theory. We give explicit tensors of low tensor rank but high monotone tensor rank, achieving a nearly optimal separation. This situation is similar to many known results in complexity theory.

1.3 The complexity of Mumford-Shaw type regularization

Computational complexity theory offers a powerful vocabulary for classifying the difficulty of computational problems, the most famous phrase of which is *NP-complete* (or *NP-hard*).

In Chapter 3, we take a natural problem in the theory of image processing and show that it is NP-hard. Specifically, there is a technique in image processing known as Mumford-Shah segmentation, for which polynomial-time solutions are known. We show that taking the slightest generalization of this technique is already NP-hard in general to solve, or even approximate up to an additive error.

Along the way, we reprove a classic result in the context of sparsity-constrained

minimization.

1.4 Full spark frames

The field of compressed sensing, roughly speaking, studies the problem of efficiently processing and reconstructing signals from under-determined information, particularly by using properties of the signal such as *sparsity* in a particular basis.

In many of the applications in compressed sensing (and especially the related field of frame theory), there is a pressing need for deterministic constructions of frames with the following property: every size- M subcollection of the M -dimensional frame elements is a spanning set. Such frames are called *full spark frames*, and in Chapter 4, we start off by providing constructions using Vandermonde matrices and discrete Fourier transforms.

Later in the chapter, we prove that full spark Parseval frames are dense in the entire set of Parseval frames, meaning full spark frames are abundant, even if one imposes a particular additional tightness constraint.

Finally, in the same vein as the previous chapter, we prove that the problem of testing whether a given matrix is full spark is hard for NP under randomized polynomial-time reductions, indicating that deterministic full spark constructions are particularly significant because they guarantee a property which is otherwise difficult to check.

1.5 Phase retrieval with polarization

In many areas of imaging science, it is difficult to measure the phase of linear measurements. As such, one often wishes to reconstruct a signal from intensity measurements, that is, perform *phase retrieval*. In Chapter 5, we provide a novel measurement design which is inspired by interferometry and exploits certain properties of expander

graphs. We also give an efficient phase retrieval procedure, and use recent results in spectral graph theory to produce a stable performance guarantee that rivals the state of the art.

We use the results of Chapter 4 freely in this chapter, as full spark frames turn out to be a quite useful concept.

1.6 Forbidden induced subgraphs of double-split graphs

The concept of *perfect graphs* is a natural concept in graph theory. In the course of proving the strong perfect graph theorem (which serves as the crowning classification result in this theory), Chudnovsky, Robertson, Seymour, and Thomas showed that every perfect graph either belongs to one of five basic classes or admits one of several decompositions. Four of the basic classes are closed under taking induced subgraphs (and have known forbidden subgraph characterizations), while the fifth one, consisting of double-split graphs, is not.

In Chapter 6, we find the forbidden induced subgraph characterization of double-split graphs. It contains 44 graphs, and so in particular is finite.

1.7 Equations resolving a conjecture of Rado on partition regularity

The field of combinatorial number theory studies many problems. Some of the most fruitful aspects of the theory have come from problem involving colorings of the integers avoiding or exhibiting particular additive or multiplicative structure.

In Chapter 7, we study the regularity of linear equations. Specifically, a linear equation L is called *k-regular* if every k -coloring of the positive integers contains

a monochromatic solution to L . Richard Rado conjectured that for every positive integer k , there exists a linear equation that is $(k - 1)$ -regular but not k -regular. We prove this conjecture by giving explicit equations with this property for every k .

This conjecture is part of problem E14 in Richard K. Guy’s book “Unsolved problems in number theory”, where it is attributed to Rado’s 1933 thesis, “Studien zur Kombinatorik”, so this problem had been unsolved for 75 years.

1.8 Publications

All of the chapters included in this work have either appeared in print or have been submitted to a journal (in modified form). They are also all available on the arXiv.org e-Print archive.

The research in Chapter 2 was performed with the help of Michael Forbes and Jacob Tsimerman. The associated article was published in the IEEE Conference on Computational Complexity [AFT11] and has not yet appeared in a journal. (This situation is not uncommon in theoretical computer science).

The research in Chapter 3 was performed with Rachel Ward. The associated article was published in IEEE Transactions on Image Processing [AW10].

The research in Chapters 4 and 5 was performed with different subsets of Afonso Bandeira, Jameson Cahill, Matthew Fickus, and Dustin G. Mixon. One associated article was published in the Journal of Fourier Analysis and Applications [ACM12]. The other has been submitted to the SIAM Journal on Imaging Sciences [ABFM12].

The research in Chapter 6 was performed with Alexandra Fradkin and Ilhee Kim. The associated article was published in the SIAM Journal on Discrete Mathematics [AFK12].

The research in Chapter 7 was performed with Jacob Tsimerman. The associated article was published in the Journal of Combinatorial Theory, Series A [AT10].

1.9 Original authorship

Like many mathematicians, I collaborate frequently. Given the context, however, it is important to understand the individual contributions to all of the research contained within this thesis.

Note that no coauthor contributed to more than one of the many chapters of this thesis, except Dustin Mixon and Jacob Tsimmerman who contributed to two each.

The research underlying Chapters 2 and 7 (particularly the latter) was performed in such a way that it is essentially impossible to isolate the individual contributions of the authors. All of the key ideas came in intense brainstorming sessions I had with Jacob Tsimmerman.

I provided the key idea and techniques of Chapter 3, a reduction from the NP-hard problem SUBSET-SUM to the problem under analysis. This reduction was relevant to independent research of Rachel Ward, which is not featured within. I similarly provided the complexity-theoretic content of Chapter 4. The rest is joint work.

I provided the key idea and techniques of Chapter 5, which is the use of expander graphs for phase retrieval.

I wrote a computer program to discover the main result of Chapter 6, which was then proven over many sessions together with Alexandra Fradkin and Ilhee Kim.

1.10 Further directions

Many of the results in this thesis naturally suggest further research. In some cases, these further avenues have already been explored.

The results of Chapter 2 have been pushed even further, to a record-holding rank of approximately $3.52n$ for an $[n] \times [n] \times [n]$ tensor. (The previous record was $3n$.) This result is not yet published and should appear in a journal paper at some point [AFT13].

Meanwhile, our results have been cited many times, of which we mention two particularly productive uses. J. M. Landsberg analyzed our tensors from the point of view of *border rank* and showed that our tensors had the minimal rank of any concise tensor, despite their having non-negligible rank. [Lan13]

Yang D. Li used our separation results involving monotone tensor rank to prove further separation results in algebraic complexity. [Li11]

Chapter 4 has, in a sense, served as a standard reference for results in compressed sensing that were previously suspected but never published. (The paper has had at least 25 citations in two years.)

Chapter 5 is part of a large group of papers pushing phase retrieval. In particular, the ideas of the paper have already found use in *sparse* retrieval problems, as well as for masked signals.

The other Chapters have been cited a few times already, but not used directly in further arguments.

Chapter 2

Tensor rank: some lower and upper bounds

2.1 Introduction

Most real-world computing treats data as Boolean, and thus made of bits. However, for some computational problems this viewpoint does not align with algorithm design. For example, the determinant is a polynomial, and computing it typically does not require knowledge of the underlying bit representation and rather treats the inputs as numbers in some field. In such settings, it is natural to consider the computation of the determinant as computing a polynomial over the underlying field, as opposed to computing a Boolean function.

When computing polynomials, just as when computing Boolean functions, there are many different models of computation to choose. The most general is the *algebraic circuit* model. Specifically, to compute a polynomial f over a field $\mathbb{F}$ in variables $x_1, \dots, x_n$, one defines a directed acyclic graph, with exactly n source nodes (each labeled with a distinct variable), a single sink node (which is thought of as the output), and internal nodes labeled with either $+$, meaning addition, or $\times$, meaning mul-

multiplication. Further, each non-leaf is restricted to have at most two children nodes. Computation is defined in the natural way: each non-source node computes the (polynomial) function of its children according to its label, and source nodes compute the variable they are labeled with. One can also consider the *algebraic formula* model, which requires the underlying graph to be a tree. In both of these models, we define the *size* of the circuit/formula to be the total number of nodes in the graph.

Neither the algebraic circuit nor the formula model are well understood, in the sense that while it can be shown that there exist polynomials which require large circuits for their computation, no explicit¹ examples of such polynomials are known. Indeed, finding such lower bounds for explicit functions is considered one of the most difficult problems in computational complexity theory. Several lower bounds are known, such as Strassen's [Str75] result (using the result of Baur-Strassen [BS83]) that the degree n polynomial $\sum_{i=1}^n x_i^n$ requires $\Omega(n \lg n)$ size circuits. However, no super-linear size lower bounds are known for constant-degree polynomials. In the case of formulas, Kalorkoti [Kal85] proved a quadratic-size lower bound for an explicit function.

One avenue for approaching improvements for both of these models is by proving lower bounds for *tensor rank*. A tensor is a generalization of a matrix, and an order- d tensor is defined as a function $T : [n]^d \rightarrow \mathbb{F}$, where $[n]$ denotes the set $\{1, \dots, n\}$. A tensor is *rank one* if it can be factorized as $T(i_1, \dots, i_d) = \prod_{j=1}^d \vec{v}_j(i_j)$ for $\vec{v}_j \in \mathbb{F}^n$. The rank of a tensor is the minimum r such that $T = \sum_{k=1}^r S_k$ for rank one tensors S_k . It can be seen that an order-2 tensor is a matrix, and the notions of rank coincide. It can also be observed that the rank of an $[n]^d$ tensor is always at most n^{d-1} , and a counting-type argument shows that over any field there exist tensors of rank at least n^{d-1}/d . A tensor is called *explicit* if $T(i_1, \dots, i_d)$ can be computed by algebraic circuits of size at most polynomial in $\text{poly}(d \lg n)$, that is, at most polynomial in the

¹A polynomial is said to be *explicit* if the coefficient of a monomial $\vec{X}^{\vec{\alpha}}$ is computable by algebraic circuits of size at most $\text{poly}(|\vec{\alpha}|)$.

size of the input $(i_1, \dots, i_d)$. All explicit tensors in this paper will also be uniformly explicit.

Interest in tensors arise from their natural correspondence with certain polynomials. Consider the sets of variables $\{X_{i,j}\}_{i \in [n], j \in [d]}$. Given a tensor $T : [n]^d \rightarrow \mathbb{F}$, one can define the polynomial

$$f_T(\{X_{i,j}\}_{i \in [n], j \in [d]}) = \sum_{i_1, \dots, i_d \in [n]} T(i_1, \dots, i_d) \prod_{j=1}^d X_{i_j, j}$$

This connection was used in the following two results. First, Strassen [Str73] showed that

Theorem (Strassen [Str73], see also [vzG88]). *For a tensor $T : [n]^3 \rightarrow \mathbb{F}$, the circuit size complexity of f_T is $\Omega(\text{rank}(T))$.*

Thus, any super-linear lower-bound for order-3 tensor rank gives a super-linear lower bounds for general arithmetic circuits, even for the constant degree polynomials. More recently, Raz [Raz10] proved

Theorem (Raz [Raz10]). *For a family of tensors $T_n : [n]^{d(n)} \rightarrow \mathbb{F}$ with $\text{rank}(T_n) \geq n^{(1-o(1))d(n)}$ and $\omega(1) \leq d(n) \leq O(\log n / \log \log n)$, the formula-size complexity of f_{T_n} is super-polynomial.*

Thus, while Strassen's result cannot be used to prove super-quadratic circuit-size lower bounds (because of the upper bounds on order-3 tensor rank), Raz's result shows that tensor rank could be used to prove very strong lower-bounds. These results motivate a study of tensor rank as a model of computation in of itself.

2.2 Prior work

Strassen's connection between order-3 tensor rank and circuit complexity further established a close connection between tensor rank to what is known as *bilinear com-*

plexity. As several important problems, such as matrix multiplication and polynomial multiplication, are bilinear, one can study their bilinear complexity, and thus their order-3 tensor rank. We interpret various prior results in the language of tensor rank. For the matrix multiplication (which corresponds to a tensor of size $[n^2] \times [n^2] \times [n^2]$), Shpilka [Shp03] showed that the tensor rank is at least $3n^2 - o(n^2)$ over $\mathbb{F}_2$, and Bläser [Blä99] earlier showed that over any field the tensor rank is at least $2.5n^2 - \Theta(n)$. For polynomial multiplication (which corresponds to a tensor of size $[2n-1] \times [n] \times [n]$), Kaminski [Kam05] showed that the tensor rank over $\mathbb{F}_q$ is known to be $(3 + 1/\Theta(q^3))n - o(n)$ and earlier work by Brown and Dobkin [BD80] showed that over $\mathbb{F}_2$ the tensor rank is at least $3.52n$. Lower bounds for these problems seem difficult, in part because strong upper bounds exist for both matrix multiplication and polynomial multiplication.

This work attempts to prove tensor rank lower bounds for *any* explicit function, not just problems of prior interest such as matrix or polynomial multiplication. Previous work in this realm include that of Ja'Ja' [Ja'79] (see their Theorem 3.6), who used the Kronecker theory of pencils to show tensor rank lower bounds of $1.5n$ for $[n] \times [n] \times [2]$ tensors, for large fields. The work was later expanded by Sumi, Miyazaki, and Sakata [SMS09] to smaller fields. However, in these works the rank is shown to be at most $1.5n$, so seemingly cannot be pushed further.

It is also worth noting that Håstad proved [Hås89, Hås90] that determining if the tensor rank of $T : [n]^3 \rightarrow \mathbb{F}$ is at most r is **NP**-hard, for $\mathbb{F}$ finite or the rationals (the problem is also known to be within **NP** for finite $\mathbb{F}$, but not known for the rationals). Implicit in his work is a tensor rank lower bound (for explicit order-3 tensors) of $4n/3$. To the best of our knowledge, the hardness of approximating tensor rank is an open question. Part of its difficulty is that any gap-preserving reduction from **NP** to tensor rank would automatically yield lower bounds for explicit tensors.

It is also a folklore result (eg. see Raz [Raz10]) that one can reshape, or embed,

a $n^{\lfloor d/2 \rfloor} \times n^{\lfloor d/2 \rfloor}$ size matrix into a order- d tensor, thus achieving a $n^{\lfloor d/2 \rfloor}$ rank lower bound for $[n]^d$ size tensors.

2.3 Our results

Our work has several components, each studying different aspects of the tensor rank problem. We first give two new methods in proving tensor rank lower bounds. In Section 2.5, we detail the first construction, which proves the best known² tensor rank lower bound for a tensor of size $[n] \times [n] \times [n]$ (over any field). In particular, using a generalization of Gaussian elimination we prove

Theorem (Corollary 2.5.4). *Let $\mathbb{F}$ be an arbitrary field. There are explicit $\{0, 1\}$ -tensors $T_n : [n]^3 \rightarrow \mathbb{F}$ such that $\text{rank}(T_n) = 3n - \Theta(\lg n)$.*

However, our analysis of this construction is exact so no further improvements can be made. In Section 2.10, we give a different order-3 tensor construction with a $3n - \Theta(\lg n)$ rank lower bound over $\mathbb{F}_2$ that has no matching upper bound, and leave as a open question what is the correct rank. In Section 2.11, we show how to extend the order-3 tensor rank lower bounds to yield a lower bound (for odd d) of $2n^{\lfloor d/2 \rfloor} + n - \Theta(d \lg n)$ for the tensor rank of an explicit 0/1 size $[n]^d$ tensor, which is improves by a factor of 2 on the folklore reshaping lower bound of $n^{\lfloor d/2 \rfloor}$.

In Section 2.6, we explore the tensor rank of permutation tensors. For matrices, permutation matrices are all full-rank and have a tight connection with the determinant. Consequently, it is natural to conjecture that a generalization of permutation matrices, which we call permutation tensors, have high rank. In particular, using a counting lower bound for Latin squares, we show that indeed there is a order-3 permutation tensor with super-linear tensor rank (over finite fields).

²When comparing this result to those listed in the prior work, it is helpful to note the differences in size of the tensors, such as comparing $[n^2]^3$ (for matrix multiplication) to our $[n]^3$. Thus, over $\mathbb{F}_2$, we essentially match Shpilka's $3n^2 - o(n^2)$ matrix multiplication result up to low-order terms.

A natural class of permutation tensors are those constructed from groups. That is, for a finite group G we can define the *group tensor* $T_G^d : G^d \rightarrow \mathbb{F}$, which is a 0/1 tensor defined by $T_G^d(g_1, \dots, g_d) = 1$ iff $g_1 \cdots g_d = 1_G$. It seems natural to conjecture that these tensors might also have high-rank. However, using representation theory we can give a strong upper bound on the rank of any group tensor (over large fields such as $\mathbb{C}$). To prove results over any field, we use interpolation methods and field-transfer results to bound the rank of any group tensor arising from an abelian group. In particular, we have the following theorem.

Theorem (Theorem 2.6.5, and Corollary 2.6.11). *Let G be a finite group. For “large” fields $\mathbb{F}$, $\text{rank}_{\mathbb{F}}(T_G^d) \leq |G|^{d/2}$. Further, for any field $\mathbb{F}$, if G is abelian then $\text{rank}_{\mathbb{F}}(T_G^d) \leq O(|G|^{1+\lg d} \lg^{d-1} |G|)$.*

In each case under consideration, we show that group tensors have rank far from the maximal $\Theta_d(|G|^{d-1})$, and thus are not good candidates for high tensor rank for large d (which are needed for Raz’s application). We are unable to place non-trivial upper bounds on the rank of T_G^d for G non-abelian and small $\mathbb{F}$, but it seems natural to conjecture that strong upper-bounds exist given the above results. While these results do not unconditionally imply any circuit lower bounds, they elucidate differences between tensor rank and matrix rank by proving that group tensors are not a viable candidate of high-rank tensors. However, conditioned on the upper bound given in Theorem 2.6.5 being tight, we are able to give super-linear tensor rank lower bounds for explicit order-3 tensors.

Finally, in Section 2.7 we explore *monotone* tensor rank. Monotone computation exploits the idea that if a polynomial only uses positive coefficients (over an ordered field such as $\mathbb{Q}$), then one might try to compute this polynomial only using positive field elements. Previous researchers have tried, in various models, to show that such restricted computation is much more inefficient than unrestricted computation. Indeed, for general algebraic circuits Valiant [Val80] has shown that allowing negative

field elements allows for an exponential improvement in the efficiency of computing certain polynomials. We continue in this line of work. In particular, we can show the following nearly optimal separation.

Theorem (Theorem 2.7.3). *Let $\mathbb{F}$ be any ordered field. There is a explicit 0/1 tensor $T : [n]^d \rightarrow \mathbb{F}$ such that $\text{rank}_{\mathbb{F}}(T) \leq dn$, but the monotone rank of T is n^{d-1} .*

2.4 Definitions and notation

We first define tensors, and give some basic facts about them. Throughout this paper, $[n]$ shall denote the set $\{1, \dots, n\}$, $\llbracket n \rrbracket$ shall denote the set $\{0, \dots, n-1\}$, and $\lg n$ shall denote the logarithm of n base 2. Further, the notation $\llbracket E \rrbracket$ (the Iverson bracket) will often be used as an indicator variable for the event E , and can be distinguished from $\llbracket n \rrbracket$ by context.

Definition 2.4.1. A *tensor* over a field $\mathbb{F}$ is a function $T : \prod_{j=1}^d [n_j] \rightarrow \mathbb{F}$. It is said to have order d and size $(n_1, \dots, n_d)$. If all of the n_j are equal to n , then T is said to have size n^d . T is said to belong to the *tensor product space* $\otimes_{j=1}^d \mathbb{F}^{n_j}$.

In later sections, the input space of a tensor will sometimes be a group or a set $\llbracket n \rrbracket$ instead of the set $[n]$. Throughout this paper $\mathbb{F}$ shall denote a arbitrary field, the variable n (or $(n_1, \dots, n_d)$) shall be reserved for the tensor size, and d shall be reserved for the order. $\mathbb{F}_q$ will denote the field on q elements. We can now define the notion of rank for tensors.

Definition 2.4.2. A tensor $T : \prod_{j=1}^d [n_j] \rightarrow \mathbb{F}$ is *simple* if for $j \in [d]$ there are vectors $\vec{v}_j \in \mathbb{F}^{n_j}$ such that $T = \otimes_{j=1}^d \vec{v}_j$. That is, for all $i_j \in [n_j]$, $T(i_1, \dots, i_d) = \prod_{j=1}^d \vec{v}_j(i_j)$ where $\vec{v}_j(i_j)$ denotes the i_j -th coordinate of $\vec{v}_j$.

Definition 2.4.3. The *rank* of a tensor $T : \prod_{j=1}^d [n_j] \rightarrow \mathbb{F}$, is defined as the minimum

number of terms in a summation of simple tensors expressing T , that is,

$$\text{rank}_{\mathbb{F}}(T) = \min \left\{ r : T = \sum_{k=1}^r \otimes_{j=1}^d \vec{v}_{j,k}, \vec{v}_{j,k} \in \mathbb{F}^{n_j} \right\}$$

Notice that by definition, a non-zero tensor is simple iff it is of rank one.

The next definition shows how identically sized order- $(d-1)$ tensors can be combined into an order- d tensor.

Definition 2.4.4. For $T_1, \dots, T_{n_d} \in \otimes_{j=1}^{d-1} \mathbb{F}^{n_j}$ define $T = [T_1 \mid \dots \mid T_{n_d}]$ by the equation $T(i_1, \dots, i_{d-1}, i_d) = T_{i_d}(i_1, \dots, i_{d-1})$. The T_{i_j} are said to be the *layers* of T (along the d -th axis). Layers along other axes are defined analogously.

Conversely, given $T \in \otimes_{j=1}^d \mathbb{F}^{n_j}$, define *the l -th layer of T (along the d -axis)*, sometimes denoted $T_l \in \otimes_{j=1}^{d-1} \mathbb{F}^{n_j}$, to be the tensor defined by $T_l(i_1, \dots, i_{d-1}) = T(i_1, \dots, i_{d-1}, l)$.

2.5 Combinatorially-defined tensors

In this section, we construct combinatorially-defined tensors and prove linear lower bounds for their rank. To do so, we use the follow fact about tensors, which is proved in Section 2.9. For matrices, this can be seen as a statement about Gaussian elimination.

Corollary (Iterative Layer Reduction, Corollary 2.9.2). *For layers*

$$S_1, \dots, S_{n_d} \in \mathbb{F}^{n_1} \otimes \dots \otimes \mathbb{F}^{n_{d-1}}$$

with $S_1, \dots, S_m$ linearly independent (as vectors in the space $\mathbb{F}^{n_1 \cdots n_{d-1}}$), there exist

constants $c_{i,j} \in \mathbb{F}$, $i \in \{1, \dots, m\}$, $j \in \{m+1, \dots, n_d\}$, such that

$$\text{rank}([S_1 \mid \dots \mid S_{n_d}]) \geq \text{rank} \left(\left[S_{m+1} + \sum_{i=1}^m c_{i,m+1} S_i \mid \dots \mid S_{n_d} + \sum_{i=1}^m c_{i,n_d} S_i \right] \right) + m \quad (2.5.1)$$

The idea of this section is to construct tensors such that we can apply Corollary 2.9.2 as many times as possible. As mentioned in Remark 2.9.3, for a $[n]^d$ tensor, the lemma can be applied at most dn times, and thus the lower bounds can at best be dn . In general, the lemma may not be able to be applied this much because the elimination of layers zeroes out too much of the tensor. However, in this section we construct tensors (for $d = 3$) such that we can almost apply the lemma dn times. The result is that we give explicit (order 3) 0/1-tensors with tensor rank exactly $3n - \Theta(\lg n)$ over any field. To begin, we apply the above corollary twice, along two different axes, to get the following lemma.

Lemma 2.5.1. *Let $A_1, \dots, A_k$ be $n \times n$ sized $\mathbb{F}$ -matrices. Let I_m denote the $m \times m$ identity matrix, and 0_m denote the $m \times m$ zero matrix. Then,*

$$\text{rank} \left(\left[\begin{array}{cc|cc} I_n & 0_n & 0_n & 0_n \\ 0_n & I_n & A_1 & 0_n \end{array} \mid \dots \mid \begin{array}{cc|cc} 0_n & 0_n & A_k & 0_n \end{array} \right] \right) \geq \text{rank}([A_1 \mid \dots \mid A_k]) + 2n \quad (2.5.2)$$

and

$$\text{rank} \left(\left[\begin{array}{ccc|ccc} 0 & 0 & 0 & 0 & 0 & 0 \\ I_n & 0_n & 0 & 0_n & 0_n & 0 \\ 0_n & I_n & 0 & A_1 & 0_n & 0 \end{array} \mid \dots \mid \begin{array}{ccc|ccc} 0 & 0 & 0 & 0_n & 0_n & 0 \\ 0_n & 0_n & 0 & A_k & 0_n & 0 \end{array} \right] \right) \geq \text{rank}([A_1 \mid \dots \mid A_k]) + 2n \quad (2.5.3)$$

where the left-hand side of Equation 2.5.3 expresses the tensor rank of a $[2n+1] \times [2n+1] \times [k]$ -sized tensor.

Proof. Notice that the left hand sides of Equation 2.5.2 and Equation 2.5.3 are equal.

This follows from applying Corollary 2.8.9 twice (using that this corollary extends to layers along any axes, not just the d -th, by applying Lemma 2.8.8), once on the layers slicing the page vertically, and once on the layers slicing the page horizontally. Thus, it is enough to show Equation 2.5.2.

We now apply Corollary 2.9.2. First, we use it on the layers slicing the page vertically and deriving that

$$\text{rank} \left(\left[\begin{array}{c|c|c|c} I_n & 0_n & 0_n & 0_n \\ \hline 0_n & I_n & A_1 & 0_n \end{array} \right] \cdots \left[\begin{array}{c|c} 0_n & 0_n \\ \hline A_k & 0_n \end{array} \right] \right) \geq \text{rank} \left(\left[\begin{array}{c|c} I_n & 0_n \\ \hline C & A_1 \end{array} \right] \cdots \left[\begin{array}{c|c} 0_n & \\ \hline & A_k \end{array} \right] \right) + n \quad (2.5.4)$$

where C is an $n \times n$ matrix of field elements defined by the constants $c_{i,j}$ of Corollary 2.9.2. Notice that the layers being dropped in the use of this corollary must be linearly independent. However, as they are the layers of $[I_n \mid 0_n \mid \cdots \mid 0_n]$ which slice the page vertically, they have exactly one 1 in the first row³, and have 0 entries elsewhere. As their non-zero entries are in different positions, they are linearly independent. Similarly, we can apply the corollary again on the remaining layers that slice the page horizontally to see that

$$\text{rank} \left(\left[\begin{array}{c|c} I_n & 0_n \\ \hline C & A_1 \end{array} \right] \cdots \left[\begin{array}{c|c} 0_n & \\ \hline & A_k \end{array} \right] \right) \geq \text{rank}([C + C' \mid A_1 \mid \cdots \mid A_k]) + n \quad (2.5.5)$$

where C' is yet another $n \times n$ matrix of field elements produced by Corollary 2.9.2. We now invoke Corollary 2.8.9 to observe that

$$\text{rank}([C + C' \mid A_1 \mid \cdots \mid A_k]) \geq \text{rank}([A_1 \mid \cdots \mid A_k]) \quad (2.5.6)$$

Combining Equations (2.5.4), (2.5.5), and (2.5.6) yields Equation (2.5.2) and thus

³It is immaterial whether we call this a “row” or “column”, as no specific orientation of these tensors was chosen.

the claim. □

Applying this lemma recursively yields the following construction.

Definition 2.5.2. Let $H : \mathbb{N} \rightarrow \mathbb{N}$ denote the Hamming weight function. That is, $H(n)$ is the number of 1's in the binary expansion of n .

Theorem 2.5.3. For $i \in \{0, \dots, \lfloor \lg n \rfloor\}$, let $S_{n,i}$ be an $n \times n$ matrix defined in the following recursive manner.

- $S_{1,0} = [1]$
- For $2n > 1$,

$$S_{2n,i} = \begin{cases} \begin{bmatrix} 0_n & 0_n \\ S_{n,i} & 0_n \end{bmatrix} & \text{if } i < \lfloor \lg n \rfloor \\ \begin{bmatrix} I_n & 0_n \\ 0_n & I_n \end{bmatrix} & \text{if } i = \lfloor \lg n \rfloor \end{cases}$$

- For $2n + 1 > 1$,

$$S_{2n+1,i} = \begin{cases} \begin{bmatrix} 0 & 0 & 0 \\ 0_n & 0_n & 0 \\ S_{n,i} & 0_n & 0 \end{bmatrix} & \text{if } i < \lfloor \lg n \rfloor \\ \begin{bmatrix} 0 & 0 & 0 \\ I_n & 0_n & 0 \\ 0_n & I_n & 0 \end{bmatrix} & \text{if } i = \lfloor \lg n \rfloor \end{cases}$$

Then, denoting $T_n = [S_{n,0} \mid \dots \mid S_{n,\lfloor \lg n \rfloor}]$,

1. T_n has size $[n] \times [n] \times [\lfloor \lg n \rfloor + 1]$.

2. $\text{rank}(T_n) = 2n - 2H(n) + 1$.

3. On inputs n and $(i, j, k) \in [n] \times [n] \times [\lfloor \lg n \rfloor + 1]$, $T_n(i, j, k)$ can be computed in polynomial time. That is, in time $O(\text{polylog}(n))$.

Proof. (1): This is clear from construction.

(2): We first note that $\lfloor \lg 2n \rfloor = \lfloor \lg n + 1 \rfloor = \lfloor \lg n \rfloor + 1$. We first prove the upper bound, and then the lower bound.

To see that $\text{rank}(T_n) \leq 2n - 2H(n) + 1$ we observe that T_n has exactly this many non-zero entries. Denote this quantity r_n . We proceed by induction on the recursive definition of the $S_{n,i}$. For $n = 1$, there is clearly exactly $2 \cdot 1 - 2H(1) + 1 = 1$ non-zero entry. For $2n > 1$, $r_{2n} = r_n + 2n$ which by induction yields $r_{2n} = (2n - 2H(n) + 1) + 2n$. Observing that $H(n) = H(2n)$, we see that $r_{2n} = 2(2n) - 2H(2n) + 1$. For $2n + 1 > 1$, $r_{2n+1} = r_n + 2n$, which by induction yields $r_{2n+1} = (2n - 2H(n) + 1) + 2n$. Noticing that $H(2n + 1) = H(n) + 1$ we have that $r_{2n+1} = 2n - 2(H(2n + 1) - 1) + 1 + 2n = 4n + 2 - 2H(2n + 1) + 1 = 2(2n + 1) - 2H(2n + 1) + 1$. Thus, the induction hypothesis shows that $r_n = 2n - 2H(n) + 1$ for all n , and thus upper-bounding the rank by this quantity.

For the rank lower bound, we use Lemma 2.5.1 and induction on the recursive definition of the $S_{n,i}$. Clearly, $\text{rank}(T_1) \geq 1$. Then for $2n > 1$, $\text{rank}(T_{2n}) \geq \text{rank}(T_n) + 2n$, and for $2n + 1 > 1$, $\text{rank}(T_{2n+1}) \geq \text{rank}(T_n) + 2n$. These are exactly the same recurrences from the proceeding paragraph, and so they have the same solution: $\text{rank}(T_n) \geq 2n - 2H(n) + 1$.

Combining these two bounds shows that $\text{rank}(T_n) = 2n - 2H(n) + 1$.

(3): This is clear from the equations defining the $S_{n,i}$. □

Another application of Corollary 2.9.2 (along the one axis it has not yet been applied) yields the following claim.

Corollary 2.5.4. Define $S_{n,i}$ as in Theorem 2.5.3. Let $n \in \mathbb{N}$ and restrict to $n \geq 2$. Then, for $i \in [n]$ define $n \times n$ matrices $S'_{n,i}$ by

$$S'_{n,i} = \begin{cases} \begin{bmatrix} S_{n-1,i-1} & 0 \\ 0 & 0 \end{bmatrix} & \text{if } i \in [\lfloor \lg(n-1) \rfloor + 1] \\ \begin{bmatrix} 0_{n-1} & \vec{e}_{i-(\lfloor \lg(n-1) \rfloor + 1)} \\ 0 & 0 \end{bmatrix} & \text{else} \end{cases}$$

where $\vec{e}_j \in \mathbb{F}^{n-1}$ is the indicator column vector where $\vec{e}_j(k) = \llbracket j = k \rrbracket$. (Notice that $\lfloor \lg(n-1) \rfloor + 1 \leq n-1$ for all $n \geq 2$.) Then, denoting $T'_n = [S'_{n,1} \mid \cdots \mid S'_{n,n}]$,

1. T'_n has size $[n]^3$.
2. $\text{rank}(T'_n) = 3n - 2H(n-1) - \lfloor \lg(n-1) \rfloor - 2 \geq 3n - \Theta(\lg n)$.
3. On inputs n and $(i, j, k) \in [n] \times [n] \times [n]$, $T'_n(i, j, k)$ can be computed in polynomial time, that is, $O(\text{polylog}(n))$.

Proof. (1): This is clear from construction.

(2): Observe that in the construction of T'_n , the matrices $S'_{n,i}$ for $i > \lfloor \lg(n-1) \rfloor + 1$ are linearly independent. Thus, applying Corollary 2.9.2, we see that

$$\text{rank}(T'_n) \geq \text{rank}([\tilde{S}'_{n,1} \mid \cdots \mid \tilde{S}'_{n,\lfloor \lg(n-1) \rfloor + 1}]) + n - (\lfloor \lg(n-1) \rfloor + 1)$$

where

$$\tilde{S}'_{n,i} = \begin{bmatrix} S_{n-1,i-1} & \vec{c}_i \\ 0 & 0 \end{bmatrix}$$

for some arbitrary vectors $\vec{c}_i \in \mathbb{F}^{n-1}$. It follows from Corollary 2.8.9 that we can drop the bottom row and last column of each of the $\tilde{S}'_{n,i}$ without increasing the rank, so

that

$$\text{rank}([\tilde{S}'_{n,1} \mid \cdots \mid \tilde{S}'_{n,\lfloor \lg(n-1) \rfloor + 1}]) \geq \text{rank}([S_{n-1,0} \mid \cdots \mid S_{n-1,\lfloor \lg(n-1) \rfloor}])$$

where the $S_{n-1,i-1}$ are as defined in Theorem 2.5.3, and as such, $\text{rank}([S_{n-1,0} \mid \cdots \mid S_{n-1,\lfloor \lg(n-1) \rfloor}]) = 2(n-1) + 2H(n-1) + 1$. Combining these inequalities yields the rank lower bound for T'_n .

(3): This is clear from the equations defining T'_n , and using the explicitness of the $S_{n-1,i-1}$ as seen from Theorem 2.5.3. $\square$

Note that this analysis is exact. Section 2.10 contains similar lower bounds over $\mathbb{F}_2$ using different methods, where no non-trivial upper bound is known.

2.6 Permutation tensors

One of the most natural families of full-rank matrices are permutation matrices. This section examines a natural generalization of permutation matrices to tensors, which we call permutation tensors. A counting argument (Proposition 2.6.2) shows that there exists order-3 permutation tensors of super-linear rank (over any fixed finite field), so it is natural to conjecture that permutation tensors may all have near-maximal rank, just as in the matrix setting. However, we show (Subsection 2.6.2) that this is false: we give tensor rank upper bounds proving that permutation tensors constructed from groups have rank much less than maximal.

We begin with the formal definition of permutation tensors.

Definition 2.6.1. Let $\mathbb{F}$ be a field, and T be a tensor $T : [n]^d \rightarrow \mathbb{F}$. T is a *permutation tensor* if T assumes only 0/1 values, and T has exactly one 1 in each generalized row. (A *generalized row*, sometimes just “row”, is the set of n inputs to T resulting from fixing $d-1$ of the coordinates, and varying the remaining coordinates).

It is not hard to see that order-2 permutation tensors and permutation matrices;

as permutation matrices are those 0/1-matrices such that each row and column have exactly one 1.

2.6.1 Rank lower bounds

We now show that there exist permutation tensors of super-linear rank (over finite fields).

Proposition 2.6.2. *Let $\mathbb{F}$ be a finite field. Then there exists a permutation tensor $T : [n]^3 \rightarrow \mathbb{F}$ of rank at least $\Omega(n \log_{|\mathbb{F}|} n)$.*

Proof. A *Latin square* is an $n \times n$ matrix, with each entry labeled with a symbol from $[n]$, such that no symbol is duplicated in any row or column. Observe that order-3 permutation tensors exactly correspond to Latin squares. We now use the following fact about Latin squares, whose proof uses lower bounds for the permanent of doubly-stochastic matrices.

Theorem ([vLW01]). *The number of $n \times n$ Latin squares is at least $(n!)^{2n}/n^{n^2}$.*

A standard counting argument completes the claim. $\square$

It remains unclear if this result generalizes to higher orders. That is, can one show that for any $k > 3$ there exist permutation tensors of rank at least $\omega(n^{\lfloor d/2 \rfloor})$?

2.6.2 Rank upper bounds

In this section we define a class of permutation tensors constructed from finite groups, and show that these tensors have rank far from maximal. We will give two rank upper-bound methods. The first method uses representation theory and accordingly only works where the group has a complete set of irreducible representations (which usually means “large” fields). The second method is based on polynomial interpolation, and while it gives worse upper bounds and only works for finite abelian groups, it gives

results over any field. Neither of these methods applies to all finite non-abelian groups over small fields, and the rank of the corresponding tensors is unclear.

Definition 2.6.3. Let G be a finite group (written multiplicatively, with identity 1_G), and $\mathbb{F}$ a field. Define the order- k group tensor $T_G^d : G^d \rightarrow \mathbb{F}$ by

$$T_G^d(g_1, \dots, g_d) = \llbracket g_1 \cdots g_d = 1_G \rrbracket$$

We first explore the representation-theory based upper bound. To do so, we first cite relevant facts from representation theory.

Theorem 2.6.4 ([Ser77]). *Let G be a finite group and $\mathbb{F}$ a field. A representation of G is a homomorphism $\rho : G \rightarrow \mathbb{F}^{d \times d}$, where d is the dimension of the representation and is denoted $\dim \rho$. The character of a representation ρ is a map $\chi_\rho : G \rightarrow \mathbb{F}$ defined by $\text{tr} \circ \rho$, that is, taking the trace of the resulting matrix of the representation.*

If $\text{char}(\mathbb{F})$ is coprime to $|G|$, and $\mathbb{F}$ contains N -th roots of unity, for N equal to the least common multiple of all of the orders of elements of G , then there exists a complete set of irreducible representations. In particular, for c denoting the number of conjugacy classes of G , there is a set of representations $\rho_1, \dots, \rho_c$ and associated characters such that (among other properties) we have

1. $\frac{1}{|G|} \sum_{i=1}^c (\dim \rho_i) \cdot \chi_i(g) = \llbracket g = e \rrbracket$
2. $\sum_{i=1}^c (\dim \rho_i)^2 = n$
3. $\dim \rho_i$ divides $|G|$

In particular, for finite abelian groups, $c = n$ and $\dim \rho_i = 1$ for all i .

Notice that property (1) in the above theorem is an instance of the column orthonormality relations of character tables, which follow from the more commonly mentioned row orthonormality relations. We now use these facts to derive upper bounds on the rank of T_G^k when the conditions to the above theorem hold.

Theorem 2.6.5. *Let G be a finite group, $d \geq 2$ and $\mathbb{F}$ a field, such that $\text{char}(\mathbb{F})$ is coprime to $|G|$, and $\mathbb{F}$ contains N -th roots of unity, for N equal to the least common multiple of the orders of elements of G . Then given the irreducible representations $\rho_1, \dots, \rho_c$ for G over $\mathbb{F}$, the order- d group tensor has $|G| \leq \text{rank}_{\mathbb{F}}(T_G^d) \leq \sum_{i=1}^c (\dim \rho_i)^d \leq |G|^{d/2}$.*

In particular, for finite abelian groups, $\text{rank}_{\mathbb{F}}(T_G^k) = |G|$.

Proof. $\text{rank}_{\mathbb{F}}(T_G^d) \geq |G|$: This follows from observing that for fixed $g_3, \dots, g_d$,

$$T_G^d(\cdot, \cdot, g_3, \dots, g_d)$$

is a permutation matrix, and thus its rank (of $|G|$) lower bounds the rank of T_G^d (over any field). This can also be seen by induction on Corollary 2.8.9.

G abelian $\implies \text{rank}_{\mathbb{F}}(T_G^k) \leq |G|$: Theorem 2.6.4 further implies $\dim \rho_i = 1$ for all irreducible representations of finite abelian groups, which implies $\text{rank}_{\mathbb{F}}(T_G^d) \leq |G|$ for abelian groups.

$\sum_{i=1}^c (\dim \rho_i)^d \leq |G|^{d/2}$: Theorem 2.6.4(2) shows that $\sum_{i=1}^c (\dim \rho_i)^2 = n$. Thus the claim is equivalent to showing that for $d \in \mathbb{Z}$, $d \geq 2$ and $n_i \in \mathbb{R}_{\geq 0}$, $\sum n_i = n \implies \sum n_i^{d/2} \leq n^{d/2}$. To show this, we first show that $(n + m)^{d/2} + 0^{d/2} = (n + m)^{d/2} \geq n^{d/2} + m^{d/2}$. To see, this, observe that assuming without loss of generality that $n \geq m$, we have that

$$(n + m)^d \geq n^d + \binom{d}{\lceil d/2 \rceil} n^{\lceil d/2 \rceil} m^{\lfloor d/2 \rfloor} + m^d \geq n^d + 2n^{d/2} m^{d/2} + m^d = (n^{d/2} + m^{d/2})^2$$

where we use that $\binom{d}{\lceil d/2 \rceil} \geq d \geq 2$. Taking square roots yields $(n + m)^{d/2} + 0^{d/2} \geq n^{d/2} + m^{d/2}$. Thus, given non-negative n_i summing to n , one can iteratively zero out certain n_i while increasing the sum $\sum n_i^{d/2}$, until only $n_1 = n$ and thus $\sum n_i^{d/2} = n^{d/2}$. Thus, this is a bound on the initial sum of $\sum n_i^{d/2}$.

$\text{rank}_{\mathbb{F}}(T_G^d) \leq \sum_{i=1}^c (\dim \rho_i)^d$: The result will follow by constructing, for each i , the

order- d tensor

$$T_{\rho_i}^d(g_1, \dots, g_d) = \chi_i(g_1 \cdots g_d)$$

in rank $(\dim \rho_i)^d$. Theorem 2.6.4(1) shows that $T_G^d = \frac{1}{|G|} \sum_{i=1}^c (\dim \rho_i) \cdot T_{\rho_i}^d$ and so distributing the $\dim \rho_i / |G|$ term inside the simple tensors yields the result (where we crucially use the restriction on the field characteristic).

Thus, all that remains is to show that $\text{rank}_{\mathbb{F}}(T_{\rho_i}^d) \leq (\dim \rho_i)^d$. Using the group homomorphism properties of the representations and expanding the definition of the trace through the matrix multiplication we see

$$T_{\rho_i}^d(g_1, \dots, g_d) = \sum_{k_1=1}^{\dim \rho_i} \cdots \sum_{k_d=1}^{\dim \rho_i} (\rho_i(g_1))_{k_1, k_2} \cdots (\rho_i(g_{d-1}))_{k_{d-1}, k_d} (\rho_i(g_d))_{k_d, k_1}$$

and one can observe that for fixed $k_1, \dots, k_d$, the function

$$(\rho_i(g_1))_{k_1, k_2} \cdots (\rho_i(g_{d-1}))_{k_{d-1}, k_d} (\rho_i(g_d))_{k_d, k_1}$$

is a simple tensor so the above shows $\text{rank}_{\mathbb{F}}(T_{\rho_i}^d) \leq (\dim \rho_i)^d$ as desired. $\square$

The above result is possibly tight, motivating the question: is there a group G and irreducible representation ρ of G such that $\text{rank}_{\mathbb{F}}(T_{\rho}^d) < (\dim \rho)^d$? As the above result is tight for abelian groups, any affirmative answer to the above question would involve a non-abelian G .

Even supposing the above result was tight, one can ask what implications this gives for circuit lower bounds, especially because group tensors are explicit when the defining group operation is efficiently computable. However, applying tensor rank lower bounds to Raz's [Raz10] result requires order- d tensors of rank $n^{(1-o(1))d}$, and Theorem 2.6.5 shows that no group tensor can achieve this rank over large fields. In particular, for the purposes of tensor rank lower bounds, the lower bounds of Corollary 2.11.2 are asymptotically (in d) as good as the rank achievable by any

group tensor (over large fields).

However, if tight, Theorem 2.6.5 would yield better lower bounds than Corollary 2.11.2 for odd d . In particular, the symmetric group $\mathfrak{S}_n$ has a complete set of irreducible representations over the rationals [Ser77]. Thus, the tightness of Theorem 2.6.5 would imply a lower bound for $\text{rank}_{\mathbb{Q}}(T_{\mathfrak{S}_n}^d)$, which is an explicit tensor. To understand this lower bound, the following fact is useful.

Theorem 2.6.6 ([VK85]). *The largest dimension of an irreducible representation of $\mathfrak{S}_n$ over $\mathbb{Q}$ is of size $\sqrt{n!}/e^{\Theta(\sqrt{n})}$*

In particular, for $d = 3$, all of the above imply that

$$\text{rank}_{\mathbb{Q}}(T_{\mathfrak{S}_n}^d) \geq |\mathfrak{S}_n|^{1.5} / e^{\Theta(\sqrt{\log|\mathfrak{S}_n|})},$$

which is $\Omega(n^{1.5-\varepsilon})$, for any $\varepsilon > 0$. Then, applying Strassen’s [Str73] result would yield $\Omega(n^{1.5-\varepsilon})$ lower bounds for the (unrestricted) circuit size of explicit degree-3 polynomials (that have 0/1 coefficients). Such a conclusion would surpass the best known circuit size lower bound even for super-constant degree polynomials, which is Strassen’s [Str75] $\Omega(n \log n)$ lower bound for degree n polynomials. Thus, tightness of Theorem 2.6.5 would have interesting consequences.

Regardless of whether the result is tight, Theorem 2.6.5 only works over “large fields” in general. In particular, it does not (in general) give insight into the rank of group tensors over fixed finite fields, or even over the rationals. To take an example, the cyclic group $\mathbb{Z}_n$ requires n -th roots of unity for its irreducible representations. While Lemma 2.6.10 does show a relation between $\text{rank}_{\mathbb{Q}}(T_{\mathbb{Z}_n}^d)$ and $\text{rank}_{\mathbb{Q}[x]/\langle x^n-1 \rangle}(T_{\mathbb{Z}_n}^d)$ (where $\mathbb{Q}[x]/\langle x^n-1 \rangle$ is the field of rationals adjoined with a n -th primitive root of unity, so Theorem 2.6.5 applies) this relationship implies nothing beyond trivial rank upper bounds. Thus, to achieve rank upper bounds for group tensors over small fields we take a different approach, one using polynomial interpolation. Our result

only applies to finite abelian groups, but is able to show that have “low” rank in this regime.

Proposition 2.6.7. *Let $\mathbb{F}$ be a field with at least $d(n-1) + 1$ elements. Let $T : \llbracket n \rrbracket^d \rightarrow \mathbb{F}$ be a tensor such that*

$$T(i_1, \dots, i_d) = \sum_{m=0}^{d(n-1)} c_m \llbracket i_1 + i_2 + \dots + i_d = m \rrbracket$$

for constants $c_m \in \mathbb{F}$. Then, $\text{rank}(T) \leq d(n-1) + 1$.

Proof. Sketch. We begin by sketching the proof. The proof follows the result of Ben-Or (as reported in Shpilka-Wigderson [SW01]) on computing the symmetric polynomials efficiently over large fields. To compute a desired polynomial $f(\vec{x})$, one can introduce a new variable α and an auxiliary polynomial $P(\alpha, \vec{x})$ such that

- P is efficiently computable, of degree at most d' in α
- For some m , $f(\vec{x}) = C_{\alpha^m}(P(\alpha, \vec{x}))$. That is, f equals the coefficient of α^m in P .

To compute f on input $\vec{x}$, we can then evaluate P on $(\alpha_1, \vec{x}), \dots, (\alpha_{d'+1}, \vec{x})$ and then use interpolation to recover $C_{\alpha^m} = f(x)$.

To apply this idea to tensors, we observe the coefficients (in the variable α) of the polynomial

$$P(\alpha, \{X_j^{(i)}\}_{i,j}) := \prod_{i=1}^d \left(X_0^{(i)} + \alpha X_1^{(i)} + \alpha^2 X_2^{(i)} + \dots + \alpha^{n-1} X_{n-1}^{(i)} \right)$$

exactly correspond to the type of tensors we are trying to produce. As P is degree at most $d(n-1)$ in α , and further P is a rank one tensor in disguise, interpolation completes the result.

Details. To apply Ben-Or’s interpolation idea to tensors, we first note the connection between tensors and polynomials. Consider the space of polynomials $\mathcal{P} :=$

$\mathbb{F}[\{X_i^{(1)}\}_{i=0}^{n-1}, \dots, \{X_i^{(d)}\}_{i=0}^{n-1}]$, that is, polynomials on the variables $X_i^{(j)}$ that are set-multilinear with respect to the sets $\{X_i^{(j)}\}_{i=0}^{n-1}$. One can call such a polynomial *simple* if it can be written as

$$\prod_{j=1}^d \left(a_{j,0} X_0^{(j)} + \dots + a_{j,n-1} X_{n-1}^{(j)} \right)$$

One can then define the $\text{rank}(p)$, for $p \in \mathcal{P}$, as the least number of simple polynomials needed to sum to p . One can observe that $\mathcal{P}$ is a $[n]^d$ tensor product space, and the notions of rank coincide. In this language, we seek to upper-bound the rank of the polynomial

$$T(\{X_i^{(j)}\}_{i,j}) = \sum_{m=0}^{d(n-1)} c_m \sum_{i_1 + \dots + i_d = m} X_{i_1}^{(1)} \dots X_{i_d}^{(d)}$$

To implement Ben-Or's method, define the auxiliary polynomial P by

$$P(\alpha, \{X_j^{(i)}\}_{i,j}) := \prod_{i=1}^d \left(X_0^{(i)} + \alpha X_1^{(i)} + \alpha^2 X_2^{(i)} + \dots + \alpha^{n-1} X_{n-1}^{(i)} \right)$$

For fixed α , this polynomial is simple. When α is considered a variable, this polynomial has degree $d(n-1)$ in α . Further, the coefficient of α^m is

$$C_{\alpha^m}(P(\alpha, \{X_i^{(j)}\}_{i,j})) = \sum_{i_1 + \dots + i_d = m} X_{i_1}^{(1)} \dots X_{i_d}^{(d)}$$

which corresponds exactly to tensors of the desired form. We can now interpret the auxiliary polynomials as polynomials in $\mathbb{F}(\{X_i^{(j)}\}_{i,j})[\alpha]$, the polynomial ring in the variable α over the field of rational functions in the $X_i^{(j)}$. As $|\mathbb{F}| > d(n-1)$, we can consider the evaluations $\{P(\alpha_l, \{X(j)_i\}_{i,j})\}_{l=0}^{d(n-1)} \in \mathbb{F}[\{X_i^{(j)}\}_{i,j}]$ for distinct $\alpha_l \in \mathbb{F}$. Polynomial interpolation means that the coefficients $C_{\alpha^m}(P)$ are recoverable from linear combinations of the evaluations of P . As the $\alpha_l \in \mathbb{F}$, the (linear) evaluation map from the coefficients of P to the evaluations is defined by a $\mathbb{F}$ -matrix. Therefore, the

inverse of this map is also defined by an $\mathbb{F}$ -matrix. Specifically, there are coefficients $a_{m,l} \in \mathbb{F}$ such that

$$C_{\alpha^m}(P(\alpha, \{X_i^{(j)}\}_{i,j})) = \sum_{l=0}^{d(n-1)} a_{m,l} P(\alpha_l, \{X_i^{(j)}\}_{i,j})$$

Therefore,

$$\begin{aligned} T(\{X_i^{(j)}\}_{i,j}) &= \sum_{m=0}^{d(n-1)} c_m \sum_{l=0}^{d(n-1)} a_{m,l} P(\alpha_l, \{X_i^{(j)}\}_{i,j}) \\ &= \sum_{l=0}^{d(n-1)} \left(\sum_{m=0}^{d(n-1)} c_m a_{m,l} \right) P(\alpha_l, \{X_i^{(j)}\}_{i,j}) \end{aligned}$$

Thus, T is in the span of $d(n-1) + 1$ simple polynomials. By moving the coefficients on the simple polynomials inside the product, this shows that T is expressible as the sum of $d(n-1) + 1$ simple polynomials. Using the above connection with tensors, this shows that the rank is at most $d(n-1) + 1$. $\square$

We now turn to using Proposition 2.6.7 to upper bound the rank group tensors formed from cyclic groups.

Corollary 2.6.8. *Let $\mathbb{F}$ be a field with at least $d(n-1) + 1$ elements. Then,*

$$\text{rank}(T_{\mathbb{Z}_n}^d) \leq d(n-1) + 1.$$

Using the Structure Theorem of Abelian Groups the following can now be shown (see Corollary 2.6.11).

Corollary 2.6.9. *Let G be a finite abelian group, and $\mathbb{F}$ be a field with at least $|G|$ elements. Then $\text{rank}_{\mathbb{F}}(T_G^d) \leq |G|^{1+\lg d}$.*

While all of the results based on Proposition 2.6.7 do not require the field to have large roots of unity, they still require the field to have large size. Thus, they seemingly

do not answer the question of the rank of group tensors over small fields. However, as the next lemma shows (with proof in Section 2.12), one can transfer results over large-sized fields to small-sized fields with a minor overhead.

Lemma 2.6.10. *Let $\mathbb{K}$ a field that extends $\mathbb{F}$. Then for any tensor $T : [n]^d \rightarrow \mathbb{F}$, $\text{rank}_{\mathbb{F}}(T) \leq (\dim_{\mathbb{F}} \mathbb{K})^{d-1} \cdot \text{rank}_{\mathbb{K}}(T)$, where $\dim_{\mathbb{F}} \mathbb{K}$ is the dimension of $\mathbb{K}$ as an $\mathbb{F}$ -vector space.*

Proof. Define $m := \dim_{\mathbb{F}} \mathbb{K}$. Thus, we can identify $\mathbb{K} = \mathbb{F}^m$ as vector spaces, where we choose that $1 \in \mathbb{K}$ is the first element in the $\mathbb{F}$ -basis for $\mathbb{K}$. This gives an $\mathbb{F}$ -algebra isomorphism μ between $\mathbb{K}$ and a sub-ring M of the $m \times m$ $\mathbb{F}$ -matrices, where M is defined as the image of μ . The map μ is defined by associating $x \in \mathbb{K}$ with the matrix inducing the linear map $\mu(x) : \mathbb{F}^m \rightarrow \mathbb{F}^m$, where $\mu(x)$ is the multiplication map of x . That is, using that $\mathbb{K} = \mathbb{F}^m$ we can see that the map $y \mapsto xy$ for $y \in \mathbb{K}$ is an $\mathbb{F}$ -linear map, and thus defines $\mu(x)$ over $\mathbb{F}^m$.

That the map is injective follows from the fact that $\mu(x)$ must map $1 \in \mathbb{K}$ to $x \in \mathbb{K}$, so x is recoverable from $\mu(x)$ (and surjectivity follows by definition of M). To see the required homomorphism properties is also not difficult. As $(x+z)y = xy + zy$ for any $x, y, z \in \mathbb{K}$, this shows that $\mu(x+z) = \mu(x) + \mu(z)$ as linear maps, and thus as matrices. Similarly, as $(xz)y = x(zy)$ for any $x, y, z \in \mathbb{K}$ it must be that $\mu(xz) = \mu(x)\mu(z)$. That this map interacts linearly in $\mathbb{F}$ implies that it is an $\mathbb{F}$ -algebra homomorphism, as desired.

Now consider a tensor $T : [n]^d \rightarrow \mathbb{F}$ with simple tensor decomposition $T = \sum_{l=1}^{\text{rank}_{\mathbb{K}}(T)} \otimes_{j=1}^d \vec{a}_{j,l}$ over $\mathbb{K}$. First observe that if we define the map $\pi : \mathbb{K} \rightarrow \mathbb{F}$ defined by

$$\pi(x) = \begin{cases} x & \text{if } x \in \mathbb{F} \\ 0 & \text{else} \end{cases}$$

then $T = \sum_{l=1}^{\text{rank}_{\mathbb{K}}(T)} \pi(\otimes_{j=1}^d \vec{a}_{j,l})$. Thus for each l , $\pi(\otimes_{j=1}^d \vec{a}_{j,l})$ is a tensor $T_l : [n]^d \rightarrow \mathbb{F}$.

We now show that $\text{rank}_{\mathbb{F}}(T_l) \leq m^{d-1}$. First observe that for $x \in \mathbb{F}$, $\mu(x)$ is a diagonal matrix. In particular, because we chose $1 \in \mathbb{K}$ to be the first element in the $\mathbb{F}$ -basis for $\mathbb{K}$, for $x \in \mathbb{F}$, $\pi(x)$ is equal to the $(1, 1)$ -th entry in $\mu(x)$. Thus, it follows that $T_l(i_1, \dots, i_d) = (\mu(\vec{a}_{1,l}(i_1)) \cdots \mu(\vec{a}_{d,l}(i_d)))_{1,1}$. By expanding out the matrix multiplication we can see that T_l is expressible as

$$T_l(i_1, \dots, i_d) = \sum_{k_1=1}^m \cdots \sum_{k_{d-1}=1}^m \mu(\vec{a}_{1,l}(i_1))_{1,k_1} \cdot \mu(\vec{a}_{2,l}(i_2))_{k_1,k_2} \cdots \mu(\vec{a}_{d-1,l}(i_{d-1}))_{k_{d-2},k_{d-1}} \cdot \mu(\vec{a}_{d,l}(i_d))_{k_{d-1},1}$$

and just as in Theorem 2.6.5 we see that for fixed k_j the summands are simple $\mathbb{F}$ -tensors, and thus $\text{rank}_{\mathbb{F}}(T_l) \leq m^{d-1}$.

Using the observation that $T = \sum_{l=1}^{\text{rank}_{\mathbb{K}}(T)} T_l$ and the above bound for the $\mathbb{F}$ -rank of T_l , we then see that $\text{rank}_{\mathbb{F}}(T) \leq (\dim_{\mathbb{F}} \mathbb{K})^{d-1} \text{rank}_{\mathbb{K}}(T)$, as desired. $\square$

With this field-transfer result, we can now state rank upper bounds for group tensors (for finite abelian groups) for any field.

Corollary 2.6.11. *Let $\mathbb{F}$ be any field, and G be a finite abelian group. Then*

$$\text{rank}_{\mathbb{F}}(T_G^d) \leq |G|^{1+\lg d} \lceil \lg |G| \rceil^{d-1}.$$

In particular, if G is cyclic, then $\text{rank}_{\mathbb{F}}(T_G^d) \leq d |G| \lceil \lg |G| \rceil^{d-1}$.

Proof. Observe that using the Structure Theorem of Finite Abelian groups, we can apply Corollary 2.12.3 to $G = \prod_{i=1}^m \mathbb{Z}_{n_i}$, and using that $d(n_i - 1) + 1 \leq dn$ (as $d \geq 2$) shows that $\text{rank}_{\mathbb{F}}(T_G^d) \leq d^m \prod n_i$. As $m \leq \lg |G|$ and $|G| = \prod n_i$, the result follows. $\square$

This last result shows that for any finite abelian group, any field and any large d , the rank of the corresponding group tensor is far from possible $\Omega(n^{d-1})$. These results

do not settle the rank of group tensors for non-abelian groups over small fields, and leaves the open question whether the methods of Theorem 2.6.5 or Proposition 2.6.7 (or other methods) can resolve this case.

2.7 Monotone tensor rank

We now explore a restricted notion of tensor rank, that of monotone tensor rank. In algebraic models of computation, monotone computation requires that the underlying field is ordered, which we now define.

Definition 2.7.1. Let $\mathbb{F}$ be a field. $\mathbb{F}$ is *ordered* if there is a linear order $<$ such that

- For all $x, y, z \in \mathbb{F}$, $x < y \implies x + z < x + y$.
- For all $x, y \in \mathbb{F}$ and $z \in \mathbb{F}_{>0}$, $x < y \implies xz < yz$.

where $\mathbb{F}_{>0} = \{x \mid x \in \mathbb{F}, x > 0\}$.

Recall that every ordered field has characteristic zero, and thus is infinite.

Over ordered fields, computation of polynomials that only use positive coefficients can be done using only positive field constants, but many works (such as [Val80]) have shown that the circuit model of computation, the restriction to positive field constants in computation leads drastically worse efficiency as compared to unrestricted computation. In this section, we show that in the tensor rank model of computation, monotone computation is also much less efficient than unrestricted computation. We first define the notion of monotone tensor rank.

Definition 2.7.2. Let $\mathbb{F}$ be an ordered field. Consider a tensor $T : \prod_{i=1}^d [n_i] \rightarrow \mathbb{F}_{\geq 0}$. Define the *monotone tensor rank* of T , denoted $\text{m-rank}(T)$, to be

$$\text{m-rank}(T) = \min \left\{ r : T = \sum_{l=1}^r \vec{v}_{l,1} \otimes \cdots \otimes \vec{v}_{l,d}, \vec{v}_{l,i} \in (\mathbb{F}_{\geq 0})^{n_i} \right\}$$

We now show an essentially maximal separation between monotone tensor rank and unrestricted tensor rank, for the explicit group tensor $T_{\mathbb{Z}_n}^d$.

Theorem 2.7.3. *Let $\mathbb{F}$ be an ordered field. Consider the group tensor $T_{\mathbb{Z}_n}^d$. Then*

1. $\text{rank}_{\mathbb{F}}(T_{\mathbb{Z}_n}^d) \leq d(n-1) + 1$
2. $\text{m-rank}_{\mathbb{F}}(T_{\mathbb{Z}_n}^d) = n^{d-1}$

Proof. Sketch. The upper bounds follow from Corollary 2.6.8, and from the trivial n^{d-1} upper bound for tensor rank.

The lower bounds follows from the observation that a non-negative simple tensor “covers” non-zero entries in $T_{\mathbb{Z}_n}^d$. It is not hard to show that if a simple tensor covers at least two non-zero entries in $T_{\mathbb{Z}_n}^d$ then it places a positive weight on a zero-entry of $T_{\mathbb{Z}_n}^d$. As this cannot be canceled out in a monotone computation, each simple tensor must cover at most one non-zero entry. As there are n^{d-1} such entries, the result follows.

Details. $\text{m-rank}_{\mathbb{F}}(T_{\mathbb{Z}_n}^d) \geq n^{d-1}$: We remark that the following lower bound will only rely on the fact that $T_{\mathbb{Z}_n}^d$ is a permutation tensor, and no other properties.

In monotone computation, there is no cancellation of terms. Thus, in a monotone simple tensor decomposition $T = \sum_{l=1}^r T_l$, one can see that the partial sums $T_{\leq m} = \sum_{l=1}^m T_l$ successively cover more and more of the non-zero entries of T . We will show that in any monotone decomposition of $T_{\mathbb{Z}_n}^d$, at most one non-zero entry can be covered by any T_l , which implies that the monotone rank is at least the number of non-zero entries, which is n^{d-1} .

We now prove that in any monotone simple tensor decomposition

$$T_{\mathbb{Z}_n}^d = \sum_{l=1}^r \otimes_{j=1}^d \vec{a}_{j,l},$$

each simple tensor $T_l := \otimes_{j=1}^d \vec{a}_{j,l}$ can cover at most one non-zero entry of $T_{\mathbb{Z}_n}^d$. Suppose

not, for contradiction. Then there is a simple tensor T_l that covers at least two non-zero entries $(i_1, \dots, i_d)$ and $(i'_1, \dots, i'_d)$ of $T_{\mathbb{Z}_n}^d$. However, these tuples must differ in at least one index, we assume without loss of generality to be index 1, so that $i_1 \neq i'_1$. Consequently, we must have that $\vec{a}_{1,l}(i_1), \vec{a}_{1,l}(i'_1) > 0$ (as all field constants are positive in monotone computation). As $\vec{a}_{j,l}(i_j) > 0$ for $j > 1$ (as $T_l(i_1, i_2, \dots, i_d) = 1$), it must be that $T_l(i'_1, i_2, \dots, i_d) = \vec{a}_{1,l}(i'_1) \prod_{j>1}^d \vec{a}_{j,l}(i_j) > 0$. However, this is a contradiction. For now this positive number at $T_l(i'_1, i_2, \dots, i_d)$ cannot be canceled out by other simple tensors in a monotone computation and we must have $T_{\mathbb{Z}_n}^d(i'_1, i_2, \dots, i_d) = 0$ by the fact that this is a permutation tensor. Thus, it must be that each simple tensor in this monotone computation can only cover a single non-zero entry of $T_{\mathbb{Z}_n}^d$, which implies the lower bound by the above argument. $\square$

2.8 Basic facts about tensors

We now prove some relevant facts about tensors that are needed for the rest of the paper.

Lemma 2.8.1. *A $\otimes_{j=1}^d \mathbb{F}^{n_j}$ -tensor is an $\prod_{j=1}^d n_j$ dimensional $\mathbb{F}$ -vector space, with standard basis $\{\otimes_{j=1}^d \vec{e}_{i_j,j}\}_{i_j \in [n_j]}$ where $\{e_{i_j,j}\}_{i_j \in [n_j]}$ is the standard basis for $\mathbb{F}^{n_j}$.*

Proof. Recall that the tensor product space $\otimes_{j=1}^d \mathbb{F}^{n_d}$ is the set of functions from $\prod_{j=1}^d [n_j]$ to $\mathbb{F}$. As a $\mathbb{F}$ -valued function space, it is thus an $\mathbb{F}$ -vector space. That it is $\prod_{j=1}^d n_h$ dimensional follows from the fact that this is the cardinality of the domain.

To see that the basis is as claimed, note that the function $\otimes_{j=1}^d \vec{e}_{i_j,j}$ is equal to the tensor $T(i'_1, \dots, i'_d) = \prod_{j=1}^d \llbracket i'_j = i_j \rrbracket$. It is then not hard to see that these tensors are a basis for the tensor product space. $\square$

Lemma 2.8.2 (Multilinearity of Tensor Product). *Suppose $j \in [d]$, $\vec{v}_j \in \mathbb{F}^{n_j}$, and $a, b \in \mathbb{F}$. In the tensor product space $\otimes_{j=1}^d \mathbb{F}^{n_j}$, for any $j_0 \in [d]$ and $\vec{w} \in \mathbb{F}^{n_{j_0}}$ the*

following identity holds:

$$\vec{v}_1 \otimes \cdots \otimes (a\vec{v}_{j_0} + b\vec{w}) \otimes \cdots \otimes \vec{v}_d = a(\vec{v}_1 \otimes \cdots \otimes \vec{v}_{j_0} \otimes \cdots \otimes \vec{v}_d) + b(\vec{v}_1 \otimes \cdots \otimes \vec{w} \otimes \cdots \otimes \vec{v}_d)$$

Proof. This follows directly from Definition 2.4.2. $\square$

We now use these properties to establish a class of rank-preserving maps on tensors.

Lemma 2.8.3. *For $j \in [d]$, consider linear maps $A_j : \mathbb{F}^{n_j} \rightarrow \mathbb{F}^{n'_j}$.*

1. *The A_j induce a function on simple tensors $\otimes_{j=1}^d \vec{v}_j \mapsto \otimes_{j=1}^d A_j \vec{v}_j$ which uniquely extends to a linear map on the tensor product spaces which is denoted $\otimes_{j=1}^d A_j : \otimes_{j=1}^d \mathbb{F}^{n_j} \rightarrow \otimes_{j=1}^d \mathbb{F}^{n'_j}$.*
2. *If the A_j are invertible, then so is $\otimes_{j=1}^d A_j$ and its inverse is given by $\otimes_{j=1}^d A_j^{-1}$.*
3. *For $T : \prod_{j=1}^d [n_j]^d \rightarrow \mathbb{F}$, $\text{rank}(T) \geq \text{rank}((\otimes_{j=1}^d A_j)(T))$, with equality if the A_j are invertible.*

Proof. (1): By Lemma 2.8.1 the tensor product space $\otimes_{j=1}^d \mathbb{F}^{n_j}$ has a basis consisting entirely of simple tensors. Thus by standard linear algebra, the map $\otimes_{j=1}^d \vec{v}_j \mapsto \otimes_{j=1}^d A_j \vec{v}_j$ on this basis extends uniquely to a linear map $\otimes_{j=1}^d A_j$ on the entire tensor product space.

It must also be shown that the map $\otimes_{j=1}^d A_j$ induced from the basis elements is also compatible with the map $\otimes_{j=1}^d \vec{v}_j \mapsto \otimes_{j=1}^d A_j \vec{v}_j$ defined on the rest of the simple tensors. This fact follows from the linearity of the A_j and the multilinearity of the tensor product, Lemma 2.8.2. That is, we first use that each $\vec{v}_j$ can be expressed in terms of the basis elements $\vec{v}_j = \sum_{i_j=1}^{n_j} c_{i_j,j} \vec{e}_{i_j,j}$ and then notice that by multilinearity

of the tensor product we have

$$\begin{aligned}
\bigotimes_{j=1}^d A_j \vec{v}_j &= \bigotimes_{j=1}^d A_j \left(\sum_{i_j=1}^{n_j} c_{i_j,j} \vec{e}_{i_j,j} \right) = \bigotimes_{j=1}^d \left(\sum_{i_j=1}^{n_j} c_{i_j,j} A_j \vec{e}_{i_j,j} \right) \\
&= \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} c_{i_1,1} \cdots c_{i_d,d} \bigotimes_{j=1}^d A_j \vec{e}_{i_j,j} \\
&= \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} c_{i_1,1} \cdots c_{i_d,d} \bigotimes_{j=1}^d A_j (\bigotimes_{j=1}^d \vec{e}_{i_j,j})
\end{aligned}$$

We observe similarly that

$$\bigotimes_{j=1}^d \vec{v}_j = \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} c_{i_1,1} \cdots c_{i_d,d} (\bigotimes_{j=1}^d \vec{e}_{i_j,j}).$$

As the unique linear map induced above *defines* $\bigotimes_{j=1}^d A_j (\bigotimes_{j=1}^d \vec{v}_j)$ as

$$\sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} c_{i_1,1} \cdots c_{i_d,d} \bigotimes_{j=1}^d A_j (\bigotimes_{j=1}^d \vec{e}_{i_j,j}),$$

this shows that $\bigotimes_{j=1}^d A_j \vec{v}_j = \bigotimes_{j=1}^d A_j (\bigotimes_{j=1}^d \vec{v}_j)$, and so the two maps agree on the simple tensors.

It should also be noted that this argument is independent of the basis chosen, as long as the basis is chosen among the simple tensors. This fact follows from the fact that the induced map on the entire space agrees with the map only defined on the simple tensors. Thus, the map $\bigotimes_{j=1}^d A_j$ is well-defined.

(2): Denote the linear maps $A := \bigotimes_{j=1}^d A_j$, and $A^{-1} := \bigotimes_{j=1}^d A_j^{-1}$. Part 1 of this lemma shows that the maps A and A^{-1} compose, in either order, to be the identity on the simple tensors. As there is a basis among the simple tensors, by Lemma 2.8.1, this means that $A \circ A^{-1}$ and $A^{-1} \circ A$ are both identity maps. Thus A^{-1} is indeed the inverse map of A .

(3): Consider a minimal simple tensor decomposition of T , so that

$$T = \sum_{l=1}^r \otimes_{j=1}^d \vec{v}_{j,l}.$$

By part 1 of this lemma, we have a simple tensor decomposition

$$(\otimes_{j=1}^d A_j)T = \sum_{l=1}^r \otimes_{j=1}^d A_j \vec{v}_{j,l}.$$

This establishes the desired rank inequality. To establish equality when the A_j are invertible it is enough to run the inequality in the opposite direction using the linear map $\otimes_{j=1}^d A_j^{-1}$ and using part 2 of this lemma. $\square$

We now use these rank-preserving maps to establish facts about tensors and their layers.

Lemma 2.8.4. *Consider $T = \otimes_{j=1}^d \vec{v}_j \in \bigotimes_{j=1}^d \mathbb{F}^{n_j}$, where T is split into layers as $T = [T_1 \mid \cdots \mid T_{n_d}]$. Then $T_l = (\vec{v}_1 \otimes \cdots \otimes \vec{v}_{d-1}) \cdot \vec{v}_d(l) \in \bigotimes_{j=1}^{d-1} \mathbb{F}^{n_j}$.*

Proof. Definition 2.4.4 and Definition 2.4.2 show that

$$T_l(i_1, \dots, i_{d-1}) := T(i_1, \dots, i_{d-1}, l) = \vec{v}_1(i_1) \cdots \vec{v}_{d-1}(i_{d-1}) \cdot \vec{v}_d(l).$$

We can then note that this is exactly the function $\vec{v}_1 \otimes \cdots \otimes \vec{v}_{d-1}$, multiplied by the scalar $\vec{v}_d(l)$, as desired. $\square$

Lemma 2.8.5. *Consider the operation of taking the l -th layer (along the d -th axis).*

This is a linear map $L_l : \bigotimes_{j=1}^d \mathbb{F}^{n_j} \rightarrow \bigotimes_{j=1}^{d-1} \mathbb{F}^{n_j}$.

Proof. Given the tensor $T(\cdot, \dots, \cdot)$. Taking the l -th layer yields

$$T_l(\cdot, \dots, \cdot) := T(\cdot, \dots, \cdot, l).$$

Thus, the statements $T = S + R \implies T_l = S_l + R_l$, and $c \in \mathbb{F}, T = cS \implies T_l = cS_l$ hold because they are simply a restriction of the above identity. $\square$

We can now prove the main lemma of this section, on how applying linear maps interacts with the layers of a tensor.

Lemma 2.8.6. *Consider $T \in \otimes_{j=1}^d \mathbb{F}^{n_j}$. Expand T into layers, so $T = [T_1 \mid \cdots \mid T_{n_d}]$.*

Let $(a_{i,j})_{i,j} \in \mathbb{F}^{m \times n_d}$ be a matrix. Define $A : \mathbb{F}^{n_d} \rightarrow \mathbb{F}^m$ to be the linear map the matrix $(a_{i,j})_{i,j}$ induces via the standard basis. Then,

$$(I \otimes \cdots \otimes I \otimes A)(T) = \left[\sum_{i_1=1}^{n_d} a_{1,i_1} T_{i_1} \mid \cdots \mid \sum_{i_m=1}^{n_d} a_{m,i_m} T_{i_m} \right]$$

Proof. The proof is in two parts. The first part proves the claim for simple tensors, and the second part extends the claim, using the linearity shown in Lemma 2.8.5, to general case.

We first prove the claim for simple tensors. Let $T = \otimes_{j=1}^d \vec{v}_j$ be a simple tensor. Let $\{\vec{e}_{i,d}\}_{i \in [n_d]}$ be the standard basis for $\mathbb{F}^{n_d}$ and $\{\vec{e}_{i',d}\}_{i' \in [m]}$ be the standard basis for $\mathbb{F}^m$. Then by expanding out in terms of the basis elements and using multilinearity, we have

$$\begin{aligned} T &= \vec{v}_1 \otimes \cdots \otimes \vec{v}_d \\ &= \sum_{i=1}^{n_d} \vec{v}_1 \otimes \cdots \otimes \vec{v}_{d-1} \otimes (\vec{v}_d(i) \vec{e}_{i,d}) \end{aligned}$$

Denote $T' := (I \otimes \cdots \otimes I \otimes A)(T)$. So then,

$$\begin{aligned}
T' &= \sum_{i=1}^{n_d} \vec{v}_1 \otimes \cdots \otimes \vec{v}_{d-1} \otimes A(\vec{v}_d(i) \vec{e}_{i,d}) \\
&= \sum_{i=1}^{n_d} \vec{v}_1 \otimes \cdots \otimes \vec{v}_{d-1} \otimes (\vec{v}_d(i) \cdot A(\vec{e}_{i,d})) \\
&= \sum_{i=1}^{n_d} \vec{v}_1 \otimes \cdots \otimes \vec{v}_{d-1} \otimes \left(\vec{v}_d(i) \cdot \sum_{i'=1}^m a_{i',i} \vec{e}_{i',d} \right) \\
&= \sum_{i=1}^{n_d} \sum_{i'=1}^m \vec{v}_d(i) \cdot a_{i',i} \cdot (\vec{v}_1 \otimes \cdots \otimes \vec{v}_{d-1} \otimes \vec{e}_{i',d})
\end{aligned}$$

By Lemma 2.8.4 and Lemma 2.8.5, we have,

$$T'_l = \sum_{i=1}^{n_d} \sum_{i'=1}^{n'_d} \vec{v}_d(i) \cdot a_{i',i} \cdot (\vec{v}_1 \otimes \cdots \otimes \vec{v}_{d-1}) \cdot \vec{e}_{i',d}(l)$$

and using that $\vec{e}_{i',d}(l) = \llbracket i' = l \rrbracket$,

$$\begin{aligned}
&= \sum_{i=1}^{n_d} \vec{v}_d(i) \cdot a_{l,i} \cdot (\vec{v}_1 \otimes \cdots \otimes \vec{v}_{d-1}) \\
&= \sum_{i=1}^{n_d} a_{l,i} T_i
\end{aligned}$$

which establishes the claim for simple tensors.

Now let $T \in \otimes_{j=1}^d \mathbb{F}^{n_j}$ be an arbitrary tensor. Consider a simple tensor expansion $T = \sum_{k=1}^r S_k$ for $S_k = \otimes_{j=1}^d \vec{v}_{j,k}$. Denote $S_{k,l}$ to be the l -th layer of S_k . So then as the S_k are simple, we have that

$$(I \otimes \cdots \otimes I \otimes A)(S_k) = \left[\sum_{i_1=1}^{n_d} a_{1,i_1} S_{k,i_1} \mid \cdots \mid \sum_{i_m=1}^{n_d} a_{m,i_m} S_{k,i_m} \right]$$

by the above analysis. So then,

$$\begin{aligned} (I \otimes \cdots \otimes I \otimes A)(T) &= (I \otimes \cdots \otimes I \otimes A) \left(\sum_{k=1}^r S_k \right) \\ &= \sum_{k=1}^r \left[\sum_{i_1=1}^{n_d} a_{1,i_1} S_{k,i_1} \mid \cdots \mid \sum_{i_m=1}^{n_d} a_{m,i_m} S_{k,i_m} \right] \end{aligned}$$

by linearity of taking layers, Lemma 2.8.5, we get

$$\begin{aligned} &= \left[\sum_{k=1}^r \sum_{i_1=1}^{n_d} a_{1,i_1} S_{k,i_1} \mid \cdots \mid \sum_{k=1}^r \sum_{i_m=1}^{n_d} a_{m,i_m} S_{k,i_m} \right] \\ &= \left[\sum_{i_1=1}^{n_d} a_{1,i_1} \left(\sum_{k=1}^r S_{k,i_1} \right) \mid \cdots \mid \sum_{i_m=1}^{n_d} a_{m,i_m} \left(\sum_{k=1}^r S_{k,i_m} \right) \right] \\ &= \left[\sum_{i_1=1}^{n_d} a_{1,i_1} T_{i_1} \mid \cdots \mid \sum_{i_m=1}^{n_d} a_{m,i_m} T_{i_m} \right] \end{aligned}$$

which is the desired result. $\square$

We now apply this to get a symmetry lemma.

Corollary 2.8.7. *Consider $T \in \otimes_{j=1}^d \mathbb{F}^{n_j}$. Expand T into layers, so $T = [T_1 \mid \cdots \mid T_{n_d}]$. For any permutation $\sigma : [n_d] \rightarrow [n_d]$,*

$$\text{rank}([T_{\sigma(1)} \mid \cdots \mid T_{\sigma(n_d)}]) = \text{rank}([T_1 \mid \cdots \mid T_{n_d}])$$

Proof. Let P be the linear transformation defined by the permutation that σ induces on the basis vectors of $\mathbb{F}^{n_d}$. Then P is invertible, and so by Lemma 2.8.3.2 the induced transformation $I \otimes \cdots \otimes I \otimes P$ is also invertible and so $\text{rank}((I \otimes \cdots \otimes I \otimes P)(T)) = \text{rank}(T)$ by Lemma 2.8.3.2. Then, by Lemma 2.8.5 we see that $(I \otimes \cdots \otimes I \otimes P)(T) = [T_1 \mid \cdots \mid T_{n_d}]$. $\square$

We also need another symmetry lemma.

Lemma 2.8.8. For $T \in \otimes_{j=1}^d \mathbb{F}^{n_j}$ and a permutation $\sigma : [d] \rightarrow [d]$, define $T' \in \otimes_{j=1}^d \mathbb{F}^{n_{\sigma(j)}}$ by $T'(i_1, \dots, i_d) = T(i_{\sigma^{-1}(1)}, \dots, i_{\sigma^{-1}(d)})$. Then, $\text{rank}(T) = \text{rank}(T')$.

Proof. We show $\text{rank}(T) \geq \text{rank}(T')$, and the equality follows by symmetry as σ is invertible. Consider a simple tensor decomposition $T = \sum_{k=1}^r \otimes_{j=1}^d \vec{v}_{j,k}$. It is then easy to see that $T' = \sum_{k=1}^r \otimes_{j=1}^d \vec{v}_{\sigma(j),k}$ by considering the equation pointwise: $T'(i_1, \dots, i_d) = T(i_{\sigma^{-1}(1)}, \dots, i_{\sigma^{-1}(d)}) = \sum_{k=1}^r \prod_{j=1}^d \vec{v}_{j,k}(i_{\sigma^{-1}(j)}) = \sum_{k=1}^r \prod_{j=1}^d \vec{v}_{\sigma(j),k}(i_j)$. The conclusion then follows by considering a minimal rank expansion. $\square$

Finally, we need a corollary about how dropping layers from a tensor affects rank.

Corollary 2.8.9. For layers $S_1, \dots, S_{n_d}, S' \in \otimes_{j=1}^{d-1} \mathbb{F}^{n_j}$, we have that

$$\text{rank}([S_1 \mid \dots \mid S_{n_d}]) \leq \text{rank}([S_1 \mid \dots \mid S_{n_d} \mid S'])$$

with equality if S' is the zero layer.

Proof. ($\leq$): The projection map P induces the map $(I \otimes \dots \otimes I \otimes P)$ which takes $[S_1 \mid \dots \mid S_{n_d} \mid S']$ to $[S_1 \mid \dots \mid S_{n_d}]$ by Lemma 2.8.5 and so Lemma 2.8.3.3 implies that the rank has not increased.

($\geq$): So now assume S' is the zero layer. Then again we apply Lemma's 2.8.5 and Lemma 2.8.3.3 but now extend the natural inclusion map $\iota : \mathbb{F}^{n_d} \rightarrow \mathbb{F}^{n_d+1}$ to a linear map $(I \otimes \dots \otimes I \otimes \iota)$ on the tensors which takes $[S_1 \mid \dots \mid S_{n_d}]$ to $[S_1 \mid \dots \mid S_{n_d} \mid 0]$, again showing that the rank has not increased. $\square$

2.9 Layer reduction

This section details a generalization of row-reduction, which we call layer-reduction. We show that layer-reduction can alter a tensor in such a way to provably reduce its rank. By showing this process can be repeated many times, a rank lower bound can be established.

The following lemma is the main technical part of this section. Håstad implicitly used⁴ a version of this lemma in his proof that tensor rank is NP-Complete [Hås89, Hås90] However, Håstad's usage requires that S_{n_d} is a rank-one tensor. This special case does not seem to directly imply our lemma, which was independently proven. While the special case is sufficient to lower-bound the combinatorially-constructed tensors of Section 2.5, the full lemma is needed to lower-bound the rank of the algebraically-constructed tensors of Section 2.10.

Lemma 2.9.1 (Layer Reduction). *For layers $S_1, \dots, S_{n_d} \in \bigotimes_{j=1}^{d-1} \mathbb{F}^{n_j}$ with S_{n_d} non-zero, there exist constants $c_1, \dots, c_{n_d-1} \in \mathbb{F}$ such that*

$$\text{rank}([S_1 \mid \dots \mid S_{n_d}]) \geq \text{rank}([S_1 + c_1 S_{n_d} \mid \dots \mid S_{n_d-1} + c_{n_d-1} S_{n_d}]) + 1$$

Proof. Denote $T := [S_1 \mid \dots \mid S_{n_d}]$. The proof is in two steps. The first step defines a linear transformation A on $\mathbb{F}^{n_d}$ such that the linear transformation $I \otimes \dots \otimes I \otimes A$ is a higher-dimensional analogue of a row-reduction step in Gaussian elimination. That is, for T' the image of T , it is seen that $T' = [S_1 + c_1 S_{n_d} \mid \dots \mid S_{n_d-1} + c_{n_d-1} S_{n_d} \mid S_{n_d}]$ by Lemma 2.8.6. The c_i are chosen in such a way so that T' has a minimal simple tensor expansion where some simple tensor R is non-zero only on the S_{n_d} -layer. In the second step, the S_{n_d} -layer is dropped and the remaining tensor $T'' = [S_1 + c_1 S_{n_d} \mid \dots \mid S_{n_d-1} + c_{n_d-1} S_{n_d}]$ no longer requires R in its simple tensor expansion and so $\text{rank}(T) \geq \text{rank}(T'') + 1$.

Consider a minimal simple tensor expansion $T = \sum_{k=1}^r \bigotimes_{j=1}^d \vec{v}_{j,k}$. Expanding the $\vec{v}_{d,k}$ in terms of basis vectors yields

$$T = \sum_{k=1}^r (\vec{v}_{1,k} \otimes \dots \otimes \vec{v}_{d-1,k}) \otimes (\vec{v}_{d,k}(1) \cdot \vec{e}_{1,d} + \dots + \vec{v}_{d,k}(n_d) \cdot \vec{e}_{n_d,d})$$

⁴Håstad's usage, and proof, is reflected by Lemmas 2, 3 and 4 (and the following discussion) of the conference version [Hås89]. The journal version [Hås90] ascribes the origin of these lemmas to Lemma 2 in the work of Hopcroft and Kerr [HK71]

and in particular Lemma 2.8.4 shows that $S_{n_d} = \sum_{k=1}^r (\vec{v}_{1,k} \otimes \cdots \otimes \vec{v}_{d-1,k}) \cdot \vec{v}_{d,k}(n_d)$. As S_{n_d} is non-zero there must be some k_0 such that $\vec{v}_{d,k_0}(n_d) \neq 0$. Define $A : \mathbb{F}^{n_d} \rightarrow \mathbb{F}^{n_d}$ to be the linear transformation defined by its action on the standard basis

$$A(\vec{e}_{i,d}) = \begin{cases} \vec{e}_{n_d,d} - \frac{\vec{v}_{d,k_0}(1)}{\vec{v}_{d,k_0}(n_d)} \vec{e}_{1,d} - \cdots - \frac{\vec{v}_{d,k_0}(n_d-1)}{\vec{v}_{d,k_0}(n_d)} \vec{e}_{n_d-1,d} & \text{if } i = n_d \\ \vec{e}_{i,d} & \text{else} \end{cases}$$

Letting I denote the identity transformation, consider the tensor $T' := (I \otimes \cdots \otimes I \otimes A)(T) \in \mathbb{F}^{n_1} \otimes \cdots \otimes \mathbb{F}^{n_d}$. By Lemma 2.8.6, we observe that $T' = [S_1 + c_1 S_{n_d} \mid \cdots \mid S_{n_d-1} + c_{n_d-1} S_{n_d} \mid S_{n_d}]$, where $c_i = -\frac{\vec{v}_{d,k_0}(i)}{\vec{v}_{d,k_0}(n_d)}$.

By Lemma 2.8.3 we have the simple tensor expansion $T' = \sum_{k=1}^r \vec{v}_{1,k} \otimes \cdots \otimes \vec{v}_{d-1,k} \otimes A\vec{v}_{d,k}$. By construction, $A(\vec{v}_{d,k_0}) = \vec{v}_{d,k_0}(n_d) \cdot \vec{e}_{n_d,d}$. Using Lemma 2.8.4 we observe that the simple tensor $\vec{v}_{1,k_0} \otimes \cdots \otimes \vec{v}_{d-1,k_0} \otimes A\vec{v}_{d,k_0}$ has non-zero entries only on the S_{n_d} -layer.

We now define the linear transformation $A' : \mathbb{F}^{n_d} \rightarrow \mathbb{F}^{n_d-1}$ defined by

$$A'(\vec{e}_{i,d}) = \begin{cases} \vec{0} & \text{if } i = n_d \\ \vec{e}_{i,d} & \text{else} \end{cases}$$

This will correspond to dropping the S_{n_d} -layer. We can compose this with A to get $A'' = A' \circ A$, defined by

$$A''(\vec{e}_{i,d}) = \begin{cases} -\frac{\vec{v}_{d,k_0}(1)}{\vec{v}_{d,k_0}(n_d)} \vec{e}_{1,d} - \cdots - \frac{\vec{v}_{d,k_0}(n_d-1)}{\vec{v}_{d,k_0}(n_d)} \vec{e}_{n_d-1,d} & \text{if } i = n_d \\ \vec{e}_{i,d} & \text{else} \end{cases}$$

So now we take $T'' = (I \otimes \cdots \otimes I \otimes A'')(T)$. By Lemma 2.8.6 we see that $T'' = [S_1 + c_1 S_{n_d} \mid \cdots \mid S_{n_d-1} + c_{n_d-1} S_{n_d}]$. Further, we observe now that by construction

$A''(\vec{v}_{d,k_0}) = \vec{0}$. This leads to the simple tensor expansion,

$$\begin{aligned}
T'' &= \sum_{k=1}^r \vec{v}_{1,k} \otimes \cdots \otimes \vec{v}_{d-1,k} \otimes A'' \vec{v}_{d,k} \\
&= \vec{v}_{1,k_0} \otimes \cdots \otimes \vec{v}_{d-1,k_0} \otimes A'' \vec{v}_{d,k_0} + \sum_{k=1, k \neq k_0}^r \vec{v}_{1,k} \otimes \cdots \otimes \vec{v}_{d-1,k} \otimes A'' \vec{v}_{d,k} \\
&= \vec{v}_{1,k_0} \otimes \cdots \otimes \vec{v}_{d-1,k_0} \otimes \vec{0} + \sum_{k=1, k \neq k_0}^r \vec{v}_{1,k} \otimes \cdots \otimes \vec{v}_{d-1,k} \otimes A'' \vec{v}_{d,k} \\
&= \sum_{k=1, k \neq k_0}^r \vec{v}_{1,k} \otimes \cdots \otimes \vec{v}_{d-1,k} \otimes A'' \vec{v}_{d,k}
\end{aligned}$$

Therefore $\text{rank}(T'') \leq r - 1 = \text{rank}(T) - 1$, and thus $\text{rank}(T) \geq \text{rank}(T'') + 1$. $\square$

The layer-reduction lemma will mostly be used via the following extension.

Corollary 2.9.2 (Iterative Layer-Reduction). *For layers $S_1, \dots, S_{n_d} \in \mathbb{F}^{n_1} \otimes \cdots \otimes \mathbb{F}^{n_{d-1}}$ with $S_1, \dots, S_m$ linearly independent (as vectors in the space $\mathbb{F}^{n_1 \cdots n_{d-1}}$), there exist constants $c_{i,j} \in \mathbb{F}$, $i \in \{1, \dots, m\}$, $j \in \{m+1, \dots, n_d\}$, such that*

$$\text{rank}([S_1 \mid \dots \mid S_{n_d}]) \geq \text{rank} \left(\left[S_{m+1} + \sum_{i=1}^m c_{i,m+1} S_i \mid \dots \mid S_{n_d} + \sum_{i=1}^m c_{i,n_d} S_i \right] \right) + m \quad (2.9.1)$$

Proof. The proof is by induction on m .

$m = 1$: This is Lemma 2.9.1, up to reordering of the layers, with the observation that the singleton set $\{S_1\}$ is linearly-independent iff S_1 is non-zero. The reordering of layers is justified by Lemma 2.8.7.

$m > 1$: By the induction hypothesis we have that

$$\text{rank}([S_1 \mid \dots \mid S_{n_d}]) \geq \text{rank} \left(\left[S_m + \sum_{i=1}^{m-1} c_{i,m} S_i \mid \dots \mid S_{n_d} + \sum_{i=1}^{m-1} c_{i,n_d} S_i \right] \right) + m - 1 \quad (2.9.2)$$

for the appropriate set of constants $c_{i,j}$. As the S_i are linearly independent, $S_m +$

$\sum_{i=1}^{m-1} c_{i,m} S_i$ is non-zero and so we can eliminate this layer from

$$\left[S_m + \sum_{i=1}^{m-1} c_{i,m} S_i \mid \dots \mid S_{n_d} + \sum_{i=1}^{m-1} c_{i,n_d} S_i \right]$$

by Lemma 2.9.1 and consequently have

$$\begin{aligned} & \text{rank} \left(\left[S_m + \sum_{i=1}^{m-1} c_{i,m} S_i \mid \dots \mid S_{n_d} + \sum_{i=1}^{m-1} c_{i,n_d} S_i \right] \right) \\ & \geq \text{rank} \left(\left[\left(S_{m+1} + \sum_{i=1}^{m-1} c_{i,m+1} S_i \right) + c_{m,m+1} \left(S_m + \sum_{i=1}^{m-1} c_{i,m} S_i \right) \mid \right. \right. \\ & \quad \left. \left. \dots \mid \left(S_{n_d} + \sum_{i=1}^{m-1} c_{i,n_d} S_i \right) + c_{m,n_d} \left(S_m + \sum_{i=1}^{m-1} c_{i,m} S_i \right) \right] \right) + 1 \end{aligned} \quad (2.9.3)$$

where the $c_{m,j}$ are new constants. Now define

$$c'_{i,j} = \begin{cases} c_{i,j} + c_{m,j} c_{i,m} & \text{if } i \neq m \\ c_{m,j} & \text{else} \end{cases} \quad (2.9.4)$$

Combining Equations (2.9.2), (2.9.3), and (2.9.4) yields the desired Equation (2.9.1). $\square$

Notice that by Lemma 2.8.8 we can in fact use Lemma 2.9.1 and Corollary 2.9.2 along any axis, not just the d -th one.

Remark 2.9.3. Lemma 2.9.1 shows that the rank of $T : \prod_{j=1}^d [n_j] \rightarrow \mathbb{F}$ is at least 1 more than the rank of some $T' : \prod_{j=1}^d [n'_j] \rightarrow \mathbb{F}$, where $n'_j = n_j$ for all $j \neq j_0$, and $n'_{j_0} = n_{j_0} - 1$. In using this lemma, the quantity $\sum_{j=1}^d n_j$ decreases by one. Therefore, we can never hope to apply this lemma more than $\sum_{j=1}^d n_j$ many times, and thus using this lemma alone will never produce lower bounds larger than this quantity. Corollary 2.9.2 simply applies Lemma 2.9.1, so the same barriers apply.

2.10 Algebraically-defined tensors

The results of this section will be field-specific, and so we no longer work over an arbitrary field.

Lemma 2.10.1. *Let $\mathbb{F}_q$ be the field of q elements. Consider $n \times n$ matrices $M_1, \dots, M_k$ over $\mathbb{F}_q$ such that all non-zero linear combinations have full-rank. Then the tensor $T = [M_1 \mid \dots \mid M_k]$ has tensor rank at least $\frac{q^k - 1}{q^k - q^{k-1}}n$.*

Proof. The proof is via the probabilistic method, using randomness to perform an analogue of gate elimination. For non-zero $\vec{c} \in \mathbb{F}_q^k$, the summation $\vec{c} \cdot \vec{M}$ will nullify terms in a simple tensor expansion with some probability. This will in expectation reduce the rank. We then invoke the hypothesis that the result is full-rank, to conclude the bound on the original rank.

Consider a minimal simple tensor decomposition $T = \sum_{i=1}^r \vec{u}_i \otimes \vec{v}_i \otimes \vec{w}_i$. For $\vec{c} \in \mathbb{F}_q^k$, consider (notation-abused) dot-product $\langle \vec{c}, \vec{M} \rangle$, which can also be written as the matrix $\sum_{i=1}^k c_i M_i$. By Lemma 2.8.6 it can be seen that this is the image of T under the linear transformation $I \otimes I \otimes A$, where A is the linear transformation that sends the basis element $\vec{e}_i$ to $c_i \vec{e}_1$. Consequently, we have that $\langle \vec{c}, \vec{M} \rangle = (I \otimes I \otimes A)(T) = \sum_{i=1}^r \vec{u}_i \otimes \vec{v}_i \otimes (A\vec{w}_i)$. Noticing that $A\vec{w}_i = \langle \vec{c}, \vec{w}_i \rangle$, and that we can then treat this as a matrix instead of one-layer tensor, we see that $\langle \vec{c}, \vec{M} \rangle = \sum_{i=1}^r \langle \vec{c}, \vec{w}_i \rangle \vec{u}_i \otimes \vec{v}_i$.

Minimality implies that $\vec{u}_i \neq 0$ for all i . So for a fixed i , the set of $\vec{c}$ such that $\langle \vec{c}, \vec{w}_i \rangle = 0$ is a 1-dimensional subspace by the Rank-Nullity theorem. Using that the field size is q , this shows that

$$\Pr_{\vec{c} \in {}_u\mathbb{F}_q^k \setminus \{\vec{0}\}} [\langle \vec{c}, \vec{w}_i \rangle \neq 0] = \frac{q^k - q^{k-1}}{q^k - 1}$$

Now define $S_{\vec{c}} := \{i \mid \langle \vec{c}, \vec{w}_i \rangle \neq 0\}$. By linearity of expectation, $\mathbb{E}_{\vec{c} \in {}_u\mathbb{F}_q^k \setminus \{\vec{0}\}} [|S_{\vec{c}}|] = \frac{q^k - q^{k-1}}{q^k - 1}r$. Thus, there exists a non-zero $\vec{c}_0$ such that $|S_{\vec{c}_0}| \leq \frac{q^k - q^{k-1}}{q^k - 1}r$. Therefore, we can write the matrix $\langle \vec{c}, \vec{M} \rangle$ as $\langle \vec{c}, \vec{M} \rangle = \sum_{i=1}^r \langle \vec{c}, \vec{w}_i \rangle \vec{u}_i \otimes \vec{v}_i = \sum_{i \in S} \langle \vec{c}, \vec{w}_i \rangle \vec{u}_i \otimes \vec{v}_i$.

The hypothesis on the M_i says that $\langle \vec{c}, \vec{M} \rangle$ is of full-rank, and therefore we have that $n \leq \text{rank}(\langle \vec{c}, \vec{M} \rangle) \leq |S| \leq \frac{q^k - q^{k-1}}{q^k - 1} r$. As $\text{rank}(T) = r$, we have that $\text{rank}(T) \geq \frac{q^k - 1}{q^k - q^{k-1}} n$. $\square$

Corollary 2.10.2. *Let $\mathbb{F}_q$ be the field of q elements. Consider $n \times n$ matrices $M_1, \dots, M_n$ over $\mathbb{F}_q$ such that all non-zero linear combinations have full-rank. Then the tensor $T = [M_1 \mid \dots \mid M_n]$ has tensor rank at least $\frac{2q-1}{q-1}n - \lceil \log_q n \rceil - \frac{q}{q-1} = \frac{2q-1}{q-1}n - \Theta(\log_q n)$.*

Proof. Let $k \leq n$ be a parameter, to be optimized over later.

Notice that the hypothesis show that the matrices M_i are linearly independent and so Corollary 2.9.2 shows that

$$\text{rank}([M_1 \mid \dots \mid M_n]) \geq \text{rank}([M_1 + M'_1 \mid \dots \mid M_k + M'_k]) + (n - k)$$

where the M'_i are linear combinations of the $M_{k+1}, \dots, M_n$. Thus, any non-zero linear combination of the $(M_i + M'_i)$ is necessarily a non-zero linear combination of the M_i . In particular, this shows that any non-zero linear combination of the $(M_i + M'_i)$ has full-rank. Thus, by Lemma 2.10.1

$$\text{rank}([M_1 + M'_1 \mid \dots \mid M_k + M'_k]) \geq \frac{q^k - 1}{q^k - q^{k-1}} n$$

and so

$$\text{rank}([M_1 \mid \dots \mid M_n]) \geq n - k + \frac{q^k - 1}{q^k - q^{k-1}} n =: f(k)$$

One can observe that $k = \log_q \left(\frac{n \ln q}{1 - 1/q} \right)$ maximizes f , but asymptotically it is

sufficient to take $k = \lceil \log_q n \rceil$. Then

$$\begin{aligned}
f(k) &= n - k + \frac{1 - q^{-k}}{1 - 1/q} n \\
&\geq n - \lceil \log_q n \rceil + \frac{1 - 1/n}{1 - 1/q} n \\
&\geq n - \lceil \log_q n \rceil + \frac{1}{1 - 1/q} n - \frac{1/n}{1 - 1/q} n \\
&\geq n - \lceil \log_q n \rceil + \frac{q}{q - 1} n - \frac{q}{q - 1} \\
&\geq \frac{2q - 1}{q - 1} n - \lceil \log_q n \rceil - \frac{q}{q - 1}
\end{aligned}$$

As $f(k)$ lower-bounds the rank by the above, this establishes the claim. $\square$

The above lemma and its corollary establish a property implying tensor rank lower bounds. We now turn to constructing tensors that have this property. Clearly we seek explicit tensors, and by this we mean that each entry of the tensor is efficiently computable.

We first observe that the property can be easily constructed given explicit field extensions of the base field $\mathbb{F}$.

Proposition 2.10.3. *Let $\mathbb{F}$ be a field and $f \in \mathbb{F}[x]$ be an irreducible polynomial of degree n . Then there exists $n \times n$ $\mathbb{F}$ -matrices $M_1, \dots, M_n$, such that all non-zero $\mathbb{F}$ -linear combinations of the M_i have full-rank. Furthermore, the entries of each matrix are computable in algebraic circuits of size $O(\text{polylog}(n) \text{poly}(|f|_0))$, where $|f|_0$ is the number of non-zero coefficients of f .*

Proof. Let $f(x) = a_n x^n + \dots + a_1 x + a_0$. Recall that $\mathbb{K} = \mathbb{F}[x]/(f)$ is a field, and because $\deg f = n$, $\mathbb{K}$ is a n -dimensional $\mathbb{F}$ -vector space, where we choose $1, x, \dots, x^{n-1}$ as the basis. This gives an $\mathbb{F}$ -algebra isomorphism μ between $\mathbb{K}$ and a sub-ring M of the $m \times m$ $\mathbb{F}$ -matrices, where M is defined as the image of μ . The map μ is defined by associating $\alpha \in \mathbb{K}$ with the matrix inducing the linear map $\mu(\alpha) : \mathbb{F}^m \rightarrow \mathbb{F}^m$, where

$\mu(\alpha)$ is the multiplication map of α . That is, using that $\mathbb{K} = \mathbb{F}^m$ we can see that the map $\beta \mapsto \alpha\beta$ for $\beta \in \mathbb{K}$ is an $\mathbb{F}$ -linear map, and thus defines $\mu(\beta)$ over $\mathbb{F}^m$.

That the map is injective follows from the fact that $\mu(\alpha)$ must map $1 \in \mathbb{K}$ to $\alpha \in \mathbb{K}$, so α is recoverable from $\mu(\alpha)$ (and surjectivity follows by definition of M). To see the required homomorphism properties is also not difficult. As $(\alpha + \gamma)\beta = \alpha\beta + \gamma\beta$ for any $\alpha, \beta, \gamma \in \mathbb{K}$, this shows that $\mu(\alpha + \gamma) = \mu(\alpha) + \mu(\gamma)$ as linear maps, and thus as matrices. Similarly, as $(\alpha\gamma)\beta = \alpha(\gamma\beta)$ for any $\alpha, \beta, \gamma \in \mathbb{K}$ it must be that $\mu(\alpha\gamma) = \mu(\alpha)\mu(\gamma)$. That this map interacts linearly in $\mathbb{F}$ implies that it is an $\mathbb{F}$ -algebra homomorphism, as desired.

In particular, this means that $\alpha \in \mathbb{K}$ is invertible iff the matrix $\mu(\alpha) \in M \subseteq \mathbb{F}^{n \times n}$ is invertible. As $\mathbb{K}$ is a field, the only non-invertible matrix in M is $\mu(0)$. The $\mathbb{F}$ -algebra homomorphism means that for $a_i \in \mathbb{F}$ and $\alpha_i \in \mathbb{K}$, the linear combination $\sum a_i \mu(\alpha_i)$ equals $\mu(\sum a_i \alpha_i)$ and so the matrix $\sum a_i \mu(\alpha_i)$ is invertible iff $\sum a_i \alpha_i \neq 0$. Thus, as $1, x, \dots, x^{n-1}$ are $\mathbb{F}$ -linearly independent in $\mathbb{K}$, it follows that the matrix $\mu(1), \mu(x), \dots, \mu(x^{n-1})$ have that all non-zero $\mathbb{F}$ -linear combinations are invertible, as desired.

We now study how to compute $\mu(x^i)$. Observe that acting as a linear map on $\mathbb{F}^n$, $\mu(x^i)$ sends $x^{i+j} \pmod{f}$. To read off the x^k component can be done with a lookup table to the coefficients of f , and thus in $O(\text{poly}(|f|_0))$ size circuits. $\square$

To make the above construction explicit, we need to show that the irreducible polynomial f can be found efficiently. We now cite the following result of Shoup [Sho90]. It says that we can find irreducible polynomials in finite fields in polynomial-time provided that the field size is fixed.

Theorem 2.10.4 ([Sho90], Theorem 4.1). *For any prime or prime power q , an irreducible polynomial of degree n in $\mathbb{F}_q[x]$ can be found in time $O(\text{poly}(nq))$.*

Combining Corollary 2.10.2, Proposition 2.10.3, and Theorem 2.10.4, we arrive at

the following result.

Corollary 2.10.5. *For any fixed prime or prime power q , over the field $\mathbb{F}_q$ there is a family of tensors T_n of size $[n]^3$ such that*

1. $\text{rank}(T_n) \geq \frac{2q-1}{q-1}n - \Theta(\log_q n)$
2. *On inputs n and $(i, j, k) \in [n]^3$, $T_n(i, j, k)$ is computable in $O(\text{poly}(nq))$*

Note that this is strictly worse than Corollary 2.5.4 in two respects. First, while this result asymptotically matches the lower bound of Corollary 2.5.4 over $\mathbb{F}_2$, the above result is only valid over finite fields, and as the field size grows, the lower bound approaches $2n - o(n)$. This seems inherent in the approach.

Further, the given construction is less explicit as computing even a single entry of the tensor might require examining all n of the coefficients in the irreducible polynomial f , preventing a $O(\text{polylog}(n))$ runtime. One method of circumventing this problem is to use *sparse irreducible polynomials*. In particular, we use the following well-known construction.

Lemma 2.10.6 ([vL99], Theorem 1.1.28). *Over $\mathbb{F}_2[x]$, the polynomial*

$$f(x) = x^{2 \cdot 3^l} + x^{3^l} + 1$$

is irreducible for any $l \geq 0$.

Observe that this allows for much faster arithmetic in the extension field, and this leads to the following result when applying the above results.

Corollary 2.10.7. *Over the field $\mathbb{F}_2$, there is a family of tensors T_n of size $[n]^3$ defined for $n = 2 \cdot 3^l$, such that*

1. $\text{rank}(T_n) \geq 3n - \Theta(\lg n)$

2. On inputs n and $(i, j, k) \in [n]^3$, $T_n(i, j, k)$ is computable in $O(\text{polylog}(n))$

Thus, this algebraic construction is also explicit, at least for some values of n . Also, this corollary is not limited to $\mathbb{F}_2$. Other constructions [GP97] are known over some other fields. However, unlike the results of Section 2.5, it is not clear if better lower bounds exist for the tensors in this section. Indeed, we do not at present know non-trivial upper bounds for the tensors given here.

2.11 Higher-order tensors

In this section we investigate order- d tensors, particularly when $d > 3$ is odd. As Raz [Raz10] shows, we can always “reshape” a lower-order tensor into a higher-order tensor without decreasing rank. Raz mentions this for reshaping an order- d tensor into an order-2 tensor (a matrix) and thus shows that there are explicit order- d tensors with rank $n^{\lfloor d/2 \rfloor}$. We use our results for order-3 tensors to derive a better bound in the case when d is odd.

We first state our reshaping lemma, keeping in mind that we again now work over an arbitrary field.

Lemma 2.11.1. *Let T be an order-3 tensor of size $[n^d] \times [n^d] \times [n]$. Then define the order- $(2d + 1)$ tensor T' of size $[n]^{2d+1}$ by*

$$T'(i_1, \dots, i_{2d+1}) = T \left(1 + \sum_{j=0}^{d-1} (i_{j+1} - 1)n^j, 1 + \sum_{j=0}^{d-1} (i_{j+(d+1)} - 1)n^j, i_{2d+1} \right)$$

Then T' has rank at least $\text{rank}(T)$.

Further, if $T(\cdot, \cdot, \cdot)$ is computable in time $f(n)$, then $T'(\cdot, \dots, \cdot)$ is computable in time $O(\text{poly}(d) \text{polylog}(n) + f(n))$.

Proof. First observe that the map $(i_1, \dots, i_d) \mapsto 1 + \sum_{j=0}^{d-1} (i_{j+1} - 1)n^j$ is a bijection from $[n]^d$ to $[n^d]$ as this is simply the base- n expansion. That this map is computable

in time $O(\text{poly}(d) \text{polylog}(n))$ establishes the claim about efficiency.

Now consider a rank r' decomposition of T'

$$T' = \sum_{l=1}^{r'} \vec{v}_{l,1} \otimes \cdots \otimes \vec{v}_{l,2d+1}$$

We now define $\vec{u}_{l,1} \in \mathbb{F}^{n^d}$. Via the bijection from above, we can write

$$\vec{u}_{l,1} \left(1 + \sum_{j=0}^{d-1} (i_{j+1} - 1)n^j \right) = \vec{v}_{l,1}(i_1) \cdots \vec{v}_{l,d}(i_d)$$

and similarly we define $\vec{u}_{l,2} \in \mathbb{F}^{n^d}$ by

$$\vec{u}_{l,2} \left(1 + \sum_{j=0}^{d-1} (i_{j+(d+1)} - 1)n^j \right) = \vec{v}_{l,d+1}(i_{d+1}) \cdots \vec{v}_{l,2d}(i_{2d})$$

, and we take $\vec{u}_{l,3} = \vec{v}_{l,2d+1} \in \mathbb{F}^n$. Thus we see that

$$T = \sum_{l=1}^{r'} \vec{u}_{l,1} \otimes \vec{u}_{l,2} \otimes \vec{u}_{l,3}$$

by examining the equation pointwise, and thus $r' \geq \text{rank}(T)$. The conclusion thus follows when taking $r' = \text{rank}(T')$. $\square$

The above lemma shows that rank lower bounds for low-order tensors extend (weakly) to rank lower bounds of higher-order tensors. We now apply this lemma to the tensor rank lower bounds of Section 2.5. It is possible to do similarly with the results of Section 2.10, but a weaker conclusion would result as those lower bounds are weaker.

Corollary 2.11.2. *For every $d \geq 1$, there is a family of $\{0,1\}$ -tensors T_n of size $[n]^{2d+1}$ such that*

$$\text{rank}(T_n) = 2n^d + n - \Theta(d \lg n).$$

Further, given n and $i_1, \dots, i_{2d+1}$, $T_n(i_1, \dots, i_{2d+1})$ is computable in polynomial time (that is, in time $O(\text{poly}(d) \text{polylog}(n))$).

Proof. We first observe that the proof of Corollary 2.5.4 extends to give a family of tensors T_n with size $[n] \times [n] \times [f(n)]$, where $\text{rank}(T_n) = 2n + f(n) - \Theta(\lg n)$, where $\omega(\lg n) \leq f(n) \leq n$. Further, these tensors have their entries computed in $O(\text{polylog}(n))$ time.

Thus, this leads to tensors of size $[n^d] \times [n^d] \times [n]$ of rank $2n^d + n - \Theta(d \lg n)$. By Lemma 2.11.1 we can then see that these tensors can be reshaped into the desired tensors, establishing the claim on the rank as well as the explicitness. $\square$

2.12 Tightness of results

Using Corollary 2.9.2, we see that Proposition 2.6.7 is essentially tight.

Proposition 2.12.1. *Let $\mathbb{F}$ be a field. Let T be a tensor $T : \llbracket n \rrbracket^d \rightarrow \mathbb{F}$ such that*

$$T(i_1, \dots, i_d) = \begin{cases} 1 & \text{if } i_1 + \dots + i_d = n \\ 0 & \text{if } i_1 + \dots + i_d > n \\ \text{unconstrained} & \text{else} \end{cases}$$

then $\text{rank}_{\mathbb{F}}(T) \geq (d-1)(n-1) + 1$.

Proof. The proof is by induction on d , using Corollary 2.9.2 to achieve a lower bound.

$d = 1$: As $T \neq 0$, its rank must be at least 1, so the result follows.

$d > 1$: Decompose T into layers along the d -th axis, so that $T = [T_0 \mid \dots \mid T_{n-1}]$. Observe that the hypothesis on T implies that for any linear combination $S = \sum_{k=0}^{n-1} c_k T_k$ it must be that $S \neq 0$. For if not, one may consider the smallest i such that $c_i \neq 0$. Then $S(n-i, 0, \dots, 0) = c_i$ by the hypothesis on T and the construction of S , which is a contradiction as $c_i \neq 0$.

Thus, Corollary 2.9.2 implies that $\text{rank}_{\mathbb{F}}(T) \geq \text{rank}_{\mathbb{F}}(T_0 + \sum_{i=1}^{n-1} a_i T_i) + (n-1)$ for some $a_i \in \mathbb{F}$. However, observing that $T' = T_0 + \sum_{i=1}^{n-1} a_i T_i$ is an order- $(d-1)$ tensor fitting the hypothesis of the induction we see that $\text{rank}_{\mathbb{F}}(T') \geq (d-2)(n-1) + 1$. Combining the above equations finishes the induction. $\square$

The above proposition shows that Proposition 2.6.7 is nearly tight, because together they show that defining $T : \llbracket n \rrbracket^d \rightarrow \mathbb{F}$ by $T(i_1, \dots, i_d) = \llbracket i_1 + \dots + i_d = n \rrbracket$, we see that $(d-1)(n-1) + 1 \leq \text{rank}_{\mathbb{F}}(T) \leq d(n-1) + 1$.

Proposition 2.6.7 was done by interpolating a univariate polynomial. By interpolating multivariate polynomials one may obtain an upper bound for the rank over group tensors arising from the direct product of cyclic groups. However, the same result is derivable in a more modular fashion, which we now present. We start with the folklore fact that tensoring two tensors multiplies their rank bounds.

Lemma 2.12.2. *Let $\mathbb{F}$ be a field. Let $T : [n]^d \rightarrow \mathbb{F}$ and $S : [m]^d \rightarrow \mathbb{F}$ be two tensors. Define $(T \otimes S) : ([n] \times [m])^d \rightarrow \mathbb{F}$ by*

$$(T \otimes S)((i_1, i'_1), \dots, (i_d, i'_d)) = T(i_1, \dots, i_d) \cdot S(i'_1, \dots, i'_d)$$

Then $\text{rank}_{\mathbb{F}}(T \otimes S) \leq \text{rank}_{\mathbb{F}}(T) \text{rank}_{\mathbb{F}}(S)$.

Proof. Suppose $T = \sum_{l=1}^r \otimes_{j=1}^d \vec{a}_{j,l}$ and $S = \sum_{l'=1}^{r'} \otimes_{j'=1}^d \vec{a}'_{j',l'}$. Then $T(i_1, \dots, i_d) = \sum_l \prod_j \vec{a}_{j,l}(i_j)$ and $S(i'_1, \dots, i'_d) = \sum_{l'} \prod_{j'} \vec{a}'_{j',l'}(i'_{j'})$. Thus, $(T \otimes S)((i_1, i'_1), \dots, (i_d, i'_d))$ equals $\sum_{l,l'} \prod_{j,j'} \vec{a}_{j,l}(i_j) \vec{a}'_{j',l'}(i'_{j'}) = \sum_{l,l'} \prod_{(j,j')} (\vec{a}_{j,l} \otimes \vec{a}'_{j',l'})_{i_j, i'_{j'}}$. Thus, as for fixed l, l' the tensor $\prod_{(j,j')} (\vec{a}_{j,l} \otimes \vec{a}'_{j',l'})_{i_j, i'_{j'}}$ is simple (as a $([n] \times [m])^d \rightarrow \mathbb{F}$ tensor), this shows the claim. $\square$

We now apply this to the direct product construction of groups.

Corollary 2.12.3. *Consider integers $n_1, \dots, n_m \in \mathbb{Z}_{\geq 2}$ and consider the finite abelian group $G = \mathbb{Z}_{n_1} \times \dots \times \mathbb{Z}_{n_m}$. Let $\mathbb{F}$ be a field with at least $\max_i (d(n_i - 1) + 1)$ elements.*

Then, $\text{rank}(T_G^d) \leq \prod_i (d(n_i - 1) + 1)$.

Proof. First observe the relevant definitions imply that $T_{\mathbb{Z}_n}^d \otimes T_{\mathbb{Z}_m}^d = T_{\mathbb{Z}_n \times \mathbb{Z}_m}^d$. Thus, the claim follows directly from Corollary 2.6.8 and Lemma 2.12.2. $\square$

We now recall the Structure Theorem of Finite Abelian Groups.

Theorem 2.12.4 (Structure Theorem of Finite Abelian Groups (see, e.g. [Art91])).

Let G be a finite abelian group. Then there are (not necessarily distinct) prime powers $n_1, \dots, n_m \in \mathbb{Z}_{\geq 2}$ such that $G = \mathbb{Z}_{n_1} \times \dots \times \mathbb{Z}_{n_m}$.

This theorem shows that Corollary 2.12.3 extends to general groups. One can get better bounds if more information is known about the group, or if results such as Theorem 2.6.5 apply, but the next result shows that even without such information group tensors from finite abelian groups have “low” rank.

Chapter 3

On the complexity of Mumford-Shah type regularization

3.1 Introduction

Consider a discrete signal $x \in \mathbb{R}^N$ sampled from a piecewise smooth signal and revealed through measurements $y = Ax + e$, where $e \in \mathbb{R}^m$ is observation noise and $A : \mathbb{R}^N \mapsto \mathbb{R}^m$ is a known linear operator identified with an $m \times N$ real matrix (representing, for instance, a blurring or partial obscuring of the data). Consider the *truncated quadratic minimization problem*,

$$\begin{aligned}\hat{x} &= \operatorname{argmin}_{x \in \mathbb{R}^N} \mathcal{J}(x), \\ \mathcal{J}(x) &= \|Ax - y\|_2^2 + \sum_{j=1}^{N-1} Q(x_{j+1} - x_j),\end{aligned}\tag{3.1.1}$$

with truncated quadratic penalty term $Q(u) = \alpha \min\{u^2, \beta\}$ parametrized by $\alpha, \beta > 0$. Since its introduction in 1984 by Geman and Geman in the context of image restoration [GG84, BZ87, Nik00], this problem has been the subject of considerable theoretical and practical interest, finding applications ranging from visual analysis to

crack detection in fracture mechanics [Ron07, Ron08]. The choice of regularization is motivated as follows: Q desires to smooth small differences $|x_{j+1} - x_j| \leq \sqrt{\beta}$ where it acts quadratically, but suspends smoothing over larger differences.

From a statistical point of view, the quadratic data-fidelity term $\|Ax - y\|_2^2$ can be viewed as a log-likelihood of the data under the hypothesis that e is Gaussian random noise, while the truncated quadratic regularization term corresponds to the energy of a piecewise Gaussian Markov random field [GG84, Bes89, JW90].

The truncated quadratic minimization problem is non-smooth and highly non-convex. However, several different characterizations of the minimizers have been unveiled [Nik00, FW10]. It is known for instance that minimizers exist and satisfy a “gap” property [Nik00, FW10]: the magnitude of successive differences of such solutions are either smaller than a first threshold or larger than a second, *strictly* larger threshold. These thresholds are independent of the observed data $y = Ax + e$ and depend only on the regularization parameters α and β . This dependence is explicit, so that a priori information about the thresholds can be incorporated into choice of regularization parameters.

When A is the $N \times N$ identity matrix, the truncated quadratic objective function can be viewed as a discretization of the *Mumford-Shah functional*¹, which motivated

¹In [Cha95], the minimizers $\hat{x} = \hat{x}_{(N)}$ of the N -dimensional truncated quadratic minimization problem with $A = I$, parameters $(\alpha_{(N)}, \beta_{(N)}) = (N^2\alpha, N\beta)$, and $y_{(N)} = y_{(j/N)}$ identified with discrete samples from a continuous function $y \in L^\infty[0, 1]$, were shown to converge to the minimizer of the Mumford-Shah functional,

$$\begin{aligned}\hat{x} &= \operatorname{argmin}_{x \in SBV[0,1]} \mathcal{F}(x), \\ \mathcal{F}(x) &= \int_{[0,1] \setminus S_x} \left((x - y)^2 + \alpha \|\nabla x\|_2^2 \right) dx + \alpha\beta |S_x|,\end{aligned}$$

over the space SBV of bounded variation functions on $[0, 1]$ with vanishing Cantor part. Note that SBV functions have a well-defined *discontinuity set* S_x of finite cardinality $|S_x|$; see [FW10] for more

the variational approach for edge detection and image segmentation with its introduction in 1988. When A is the identity matrix as such, the truncated quadratic minimization problem can be solved in polynomial-time using dynamic programming [Cha95]. However, for general $m \times N$ matrices A , existing algorithms for minimizing the functional (3.1.1) guarantee convergence to *local* minimizers at best [FW10].

In this paper, we show that the truncated quadratic minimization problem is NP-hard in general, certifying that the present convergence guarantees are the best one could hope for. Consequently, the Mumford-Shah functional (3.1.2) cannot be tractably extended to general inverse problems.

3.2 Truncated quadratic minimization reformulated

It will be helpful to recast the truncated quadratic minimization problem in terms of the discrete differences $u_j = x_{j+1} - x_j$, effectively decoupling the action of the regularization term Q . We may express this change of variables in matrix notation as $u = Dx$, with $D : \mathbb{R}^N \rightarrow \mathbb{R}^{N-1}$ the *discrete difference* matrix,

$$D = \begin{pmatrix} -1 & 1 & 0 & \dots & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 \\ \dots & & & & & \\ 0 & 0 & \dots & \dots & -1 & 1 \end{pmatrix}.$$

The null space of D , which we denote in the following by $\mathcal{N}(\mathcal{D})$, is simply the one-dimensional subspace of constant vectors in $\mathbb{R}^N$. The orthogonal projection of a vector $x \in \mathbb{R}^N$ onto this subspace is the constant vector c whose entries coincide with the mean value $\frac{1}{N} \sum_{j=1}^N x_j$ of x , while its projection onto the orthogonal complement

details.

of $\mathcal{N}(\mathcal{D})$ is given by the least squares solution $D^\dagger Dx$, where $D^\dagger$ is the pseudo-inverse matrix of D in the Moore-Penrose sense. These observations yield the orthogonal decomposition $x = D^\dagger Dx + c$, or, incorporating the substitution $u = Dx$, the decomposition $x = D^\dagger u + c$.

Minimization of $\mathcal{J}$, recast in terms of the variables u and c , becomes

$$\begin{aligned} (\hat{c}, \hat{u}) &= \operatorname{argmin}_{c \in N(D), u \in \mathbb{R}^{N-1}} \mathcal{J}(c, u), \\ \mathcal{J}(c, u) &= \|AD^\dagger u + Ac - y\|_2^2 + \sum_{j=1}^{N-1} Q(u_j), \end{aligned} \quad (3.2.1)$$

where the primal minimizer $\hat{x}$ and (c, u) -minimizer $(\hat{c}, \hat{u})$ are interchangeable according to $\hat{x} = D^\dagger \hat{u} + \hat{c}$.

If the null space of A contains the constant vectors, such as if $A = TD$ for an $m \times (N-1)$ matrix T , the minimization problem (3.2.1) reduces to a function of u only,

$$\hat{u} = \operatorname{argmin}_{u \in \mathbb{R}^{N-1}} \|AD^\dagger u - y\|_2^2 + \sum_{j=1}^{N-1} Q(u_j).$$

Making the substitution $A = TD$ and using that $DD^\dagger = I$, we see in particular that any optimization problem of the form $\hat{u} = \operatorname{argmin}_{u \in \mathbb{R}^{N-1}} \|Tu - y\|_2^2 + \sum_{j=1}^{N-1} Q(u_j)$ can be identified with an instance of a truncated quadratic minimization problem (3.1.1). To summarize,

Lemma 3.2.1. *Let $T : \mathbb{R}^{N-1} \mapsto \mathbb{R}^m$ be a linear operator identified with a matrix of $\mathbb{R}^{m \times (N-1)}$. The minimization problem*

$$\hat{u} = \operatorname{argmin}_{u \in \mathbb{R}^{N-1}} \|Tu - y\|_2^2 + \sum_{j=1}^{N-1} Q(u_j) \quad (3.2.2)$$

corresponds to a truncated quadratic minimization problem

$$\hat{x} = \operatorname{argmin}_{x \in \mathbb{R}^N} \|TDx - y\|_2^2 + \sum_{j=1}^{N-1} Q(x_{j+1} - x_j) \quad (3.2.3)$$

in the sense that $\hat{x} = D^\dagger \hat{u}$ is a minimizer for (3.2.3) if $\hat{u}$ minimizes (3.2.2), while $\hat{u} = D\hat{x}$ is a minimizer for (3.2.2) if $\hat{x}$ is a minimizer for (3.2.3).

3.3 Reduction to SUBSET-SUM

Recall that the complexity class NP consists of all problems whose solution can be verified in polynomial time given a certificate for the answer. For example, the problem SUBSET-SUM is to determine, given nonzero integers $a_1, \dots, a_k$ and C , whether or not there exists a subset S of $\{1, \dots, k\}$ such that $\sum_{i \in S} a_i = C$. This problem is in NP because given any particular subset S , we can easily check whether or not its corresponding sum is zero.

Further recall that a *polynomial-time many-one reduction* from a problem A to a problem B is an algorithm that transforms an instance of A to an instance of B with the same answer in time polynomial with respect to the number of bits used to represent the instance of A . Intuitively, this captures the notion that A is no harder than B , up to polynomial factors, and accordingly one may write $A \leq B$. Finally, a problem B is called NP-hard if every problem in NP is reducible to B . (Note that $A \leq B$ and $B \leq C$ imply $A \leq C$, so if an NP-hard problem B reduces to a problem C , then C is NP-hard as well.) NP-hard problems can not be solved in polynomial time unless $P=NP$.

In order to prove our NP-hardness result, we show that the known NP-hard problem SUBSET-SUM admits a polynomial-time reduction to an instance of the trun-

cated quadratic minimization problem. Moreover, we show that any algorithm that could efficiently *approximate* this minimum (to within an additive error) could solve SUBSET-SUM efficiently as well; that is, the search for even an approximate solution to a truncated quadratic minimization problem is NP-hard as well.

Theorem 3.3.1. *Let $a_1, \dots, a_k$ and C be given nonzero integers. Then there exists a subset S of $\{1, \dots, k\}$ such that $\sum_{i \in S} a_i = C$ if and only if $\min_{x \in \mathbb{R}^{2k}} f(x) \leq k$, where*

$$f(x) = \left(C - \sum_{i=1}^k a_i x_i \right)^2 + P \cdot \sum_{i=1}^k (1 - x_i - x_{i+k})^2 + \sum_{i=1}^{2k} \min \left(1, \frac{x_i^2}{\varepsilon^2} \right),$$

with $0 < \varepsilon \leq \frac{1}{4(\sum_i |a_i|)}$ and $P \geq \frac{2k}{\varepsilon^2}$. Moreover, this minimum is never strictly between k and $k + \frac{1}{4}$.

Proof. Call a subset $S \subseteq \{1, \dots, k\}$ *good* if $\sum_{i \in S} a_i = C$. If a good subset S exists, we may set

$$x_i = \begin{cases} 1 & \text{if } i \in S, \\ 0 & \text{if } i \notin S, \end{cases} \quad \text{and} \quad x_{i+k} = \begin{cases} 0 & \text{if } i \in S, \\ 1 & \text{if } i \notin S, \end{cases}$$

for $1 \leq i \leq k$. Then $f(x) = k$ because the first and second terms vanish, and there are exactly k nonzero x_i s. Therefore, if a good subset exists, the minimum is at most k .

Suppose no good subset exists, yet there exists x such that $f(x) < k + \frac{1}{4}$. Consider the k pairs of coordinates x_i, x_{i+k} for $1 \leq i \leq k$. If both $|x_i|$ and $|x_{i+k}|$ were less than ε , a single summand in the second term already exceeds $2k$, as $P \cdot (1 - x_i - x_{i+k})^2 \geq P \cdot (1 - 2\varepsilon)^2 \geq 2k$. Therefore, at least one of $|x_i|$ and $|x_{i+k}|$ exceeds ε and the third term is already at least k . If more than one of $|x_i|$ and $|x_{i+k}|$ exceeded ε , then the third term would be at least $k + 1$, so exactly one of the coordinates in each pair exceeds ε in absolute value. If $|x_i| \leq \varepsilon$, then $|x_{i+k} - 1| \leq 2\varepsilon$ as

otherwise $P \cdot (1 - x_i - x_{i+k})^2 \geq P \cdot (\varepsilon)^2 \geq 2k$; the symmetric holds if $|x_{i+k}| \leq \varepsilon$, so all of the x_i are within 2ε of either 0 or 1.

Let $\bar{x}_i$ be the closer of 0 and 1 to x_i . Then because no good subset exists and C and the a_i are integers, $\left|C - \sum_{i=1}^k a_i \bar{x}_i\right| \geq 1$. It follows that the first term $\left(C - \sum_{i=1}^k a_i x_i\right)^2 \geq (1 - \sum 2\varepsilon a_i)^2 \geq (\frac{1}{2})^2 = \frac{1}{4}$. But the third term was already at least k , so this is a contradiction. Therefore, if no good subset exists, we must have $\min_x f(x) \geq k + \frac{1}{4}$. $\square$

Corollary 3.3.2. *Solving the truncated quadratic regularization problem, even to within an additive error, is NP-hard.*

Proof. In light of Lemma 3.2.1, minimization of the function f is a truncated quadratic minimization problem, with $m = k + 1$ and $N = 2k$. Therefore, we have reduced the known NP-hard problem SUBSET-SUM to a truncated quadratic minimization problem. It remains to verify that this reduction is polynomial-time. To see this, note that Theorem 3.3.1 ensures that the minimum of f is either at most k or at least $k + \frac{1}{4}$; thus, we only need to approximate each of the polynomially-many entries in the matrices and vectors in f to within a number of bits that is polynomial compared to the number of bits needed to represent $\sum_i |a_i|$. $\square$

3.4 Connection to sparse recovery

The only properties of the quadratic regularization term $\min\{1, x_i^2/\varepsilon^2\}$ needed for Theorem 3.3.1 were that it be bounded between 0 and 1, equal to 0 if $x_i = 0$, and equal to 1 if $|x_i| \geq \varepsilon$. Indeed, Theorem 3.3.1 holds for *any* regularization term

satisfying these properties; for example, one could consider hard thresholding,

$$|x|_0 = \begin{cases} 0 & x = 0, \\ 1 & x \neq 0, \end{cases}$$

which generates the ℓ_0 “counting norm” $\|x\|_0 = \sum_{j=1}^N |x_j|_0$. We then reprove the known result that the ℓ_0 -regularized optimization problem,

$$\hat{u} = \operatorname{argmin}_{u \in \mathbb{R}^N} \|Tu - y\|_2^2 + \gamma \|u\|_0,$$

is NP-hard in general. This functional is of considerable interest in the emerging area of sparse recovery, as it is guaranteed to produce sparse solutions for sufficiently large γ and over a certain class of matrices². In this light, the truncated quadratic minimization problem may be interpreted as a relaxation of the ℓ_0 -regularized optimization problem, and our main result as showing that even such *relaxations* of the ℓ_0 -regularized functional are NP-hard.

²See [CW08] for matrix constructions that admit polynomial-time recovery algorithms for the ℓ_0 -regularized optimization problem. All constructions at present involve an element of randomness, and a complete characterization of such matrices forms the core of the area known as *compressed sensing*.

Chapter 4

Full spark frames

4.1 Introduction

A *frame* over a Hilbert space $\mathcal{H}$ is a collection of vectors $\{f_i\}_{i \in \mathcal{I}} \subseteq \mathcal{H}$ with *frame bounds* $0 < A \leq B < \infty$ such that

$$A\|x\|^2 \leq \sum_{i \in \mathcal{I}} |\langle x, f_i \rangle|^2 \leq B\|x\|^2 \quad \forall x \in \mathcal{H}. \quad (4.1.1)$$

For a finite frame $\{f_n\}_{n=1}^N$ in an M -dimensional Hilbert space $\mathbb{H}_M$, the upper frame bound B which satisfies (4.1.1) trivially exists. In this finite-dimensional setting, the notion of being a frame solely depends on the lower frame bound A being strictly positive, which is equivalent to having the frame elements $\{f_n\}_{n=1}^N$ span the Hilbert space.

In practice, frames are chiefly used in two ways. The *synthesis operator* $F: \mathbb{C}^N \rightarrow \mathbb{H}_M$ of a finite frame $F = \{f_n\}_{n=1}^N$ is defined by $Fy := \sum_{n=1}^N y[n]f_n$. As such, the $M \times N$ matrix representation F of the synthesis operator has the frame elements $\{f_n\}_{n=1}^N$ as columns, with the natural identification $\mathbb{H}_M \cong \mathbb{C}^M$. Note that here and throughout, we make no notational distinction between a frame F and its synthesis operator. The adjoint of the synthesis operator is called the *analysis operator* $F^*: \mathbb{H}_M \rightarrow \mathbb{C}^N$,

defined by $(F^*x)[n] := \langle x, f_n \rangle$.

Classically, frames are used to redundantly decompose signals, $y = F^*x$, before synthesizing the corresponding frame coefficients, $z = Fy = FF^*x$, and so the *frame operator* $FF^*: \mathbb{H}_M \rightarrow \mathbb{H}_M$ is often analyzed to determine how well this process preserves information about the original signal x . In particular, if the frame bounds are equal, the frame operator has the form $FF^* = AI_M$, and so signal reconstruction is rather painless: $x = \frac{1}{A}FF^*x$; in this case, the frame is called *tight*. Oftentimes, it is additionally desirable for the frame elements to have unit norm, in which case the frame is a *unit norm frame*. Moreover, the *worst-case coherence* between unit norm frame elements $\mu := \max_{i \neq j} |\langle f_i, f_j \rangle|$ satisfies $\mu^2 \geq \frac{N-M}{M(N-1)}$, and when equality is achieved, the frame is necessarily tight with $|\langle f_i, f_j \rangle| = \mu$ for all distinct pairs $i, j \in \{1, \dots, N\}$ [SH03]; in this case, the frame is called an *equiangular tight frame* (ETF).

The utility of unit norm tight frames and ETFs is commonly expressed in terms of a scenario in which frame coefficients $\{(F^*x)[n]\}_{n=1}^N$ are transmitted over a noisy or lossy channel before reconstructing the signal:

$$y = \mathcal{D}(F^*x), \quad \tilde{x} = (FF^*)^{-1}Fy, \quad (4.1.2)$$

where $\mathcal{D}(\cdot)$ represents the channel's random and not-necessarily-linear deformation process. Using an additive white Gaussian noise model, Goyal [Goy98] established that, of all unit norm frames, unit norm tight frames minimize mean squared error in reconstruction. For the case of a lossy channel, Holmes and Paulsen [HP04] established that, of all tight frames, unit norm tight frames minimize worst-case error in reconstruction after one erasure, and that equiangular tight frames minimize this error after two erasures. We note that the reconstruction process in (4.1.2), namely the application of $(FF^*)^{-1}F$, is inherently blind to the effect of the deformation process

of the channel. This contrasts with Püschel and Kovačević's more recent work [PK05], which describes an adaptive process for reconstruction after multiple erasures; we will return to this idea later.

Another application of frames is sparse signal processing. This field concerns signal classes which are sparse in some basis. As an example, natural images tend to be nearly sparse in the wavelet basis [DDEK12]. Some applications have signal sparsity in the identity basis [MQKF11]. Given a signal x , let Ψ represent its sparsifying basis and consider

$$y = F\Psi^*x + e, \tag{4.1.3}$$

where F is the $M \times N$ synthesis operator of a frame, Ψ^*x has at most K nonzero entries, and e is some sort of noise. When given measurements y , one typically wishes to reconstruct the original vector x ; this process is commonly referred to as *compressed sensing* since we often take $M \ll N$. Note that y can be viewed as a noisy linear combination of a few unknown frame elements, and the goal of compressed sensing is to determine which frame elements are active in the combination, and further estimate the scalar multiples used in this combination.

In order to have any sort of inversion process for (4.1.3), even in the noiseless case where $e = 0$, F must map K -sparse vectors injectively. That is, we need $Fx_1 \neq Fx_2$ for any distinct K -sparse vectors x_1 and x_2 . Subtraction then gives that $2K$ -sparse vectors like $x_1 - x_2$ cannot be in the nullspace of F . This identification led to the following definition [DE03]:

Definition 4.1.1. The *spark* of a matrix F is the size of the smallest linearly dependent subset of columns, i.e.,

$$\text{Spark}(F) = \min\{\|x\|_0 : Fx = 0, x \neq 0\}.$$

Using the above analysis, Donoho and Elad [DE03] showed that a signal x with

sparsity level $< \text{Spark}(F)/2$ is necessarily the unique sparsest solution to $y = Fx$, and furthermore, there exist signals x of sparsity level $\geq \text{Spark}(F)/2$ which are not the unique sparsest solution to $y = Fx$. This demonstrates that matrices with larger spark are naturally equipped to support signals with larger sparsity levels. One is naturally led to consider matrices that support the largest possible sparse signal class; we say an $M \times N$ matrix F is *full spark* if its spark is as large as possible, i.e., $\text{Spark}(F) = M + 1$. Equivalently, $M \times N$ full spark matrices have the property that every $M \times M$ submatrix is invertible; as such, a full spark matrix is necessarily full rank, and therefore a frame.

Consider a matrix whose entries are independent continuous random variables. Certainly, the matrix is full spark with probability one, but this random process does not allow one to control certain features of the matrix, such as unit-norm tightness or equiangularity. Indeed, ETFs are notoriously difficult to construct, but they appear to be particularly well-suited for sparse signal processing [BCM12, FM11, MQKF11]. For example, Bajwa et al. [BCM12] uses Steiner ETFs to recover the support of Ψ^*x in (4.1.3); given measurements y , the largest entries in the back-projection F^*y often coincide with the support of Ψ^*x . However, Steiner ETFs have particularly small spark [FMT12], and back-projection will correctly identify the support of x even when the corresponding columns in F are dependent; in this case, there is no way to estimate the nonzero entries of Ψ^*x . This illustrates one reason to build deterministic full spark frames: their submatrices are necessarily invertible, making it possible to estimate these nonzero entries.

For another application of full spark frames, we return to the problem (4.1.2) of reconstructing a signal from distorted frame coefficients. Specifically, we consider Püschel and Kovačević's work [PK05], which focuses on an Internet-like channel that is prone to multitudes of erasures. In this context, they reconstruct the signal after first identifying which frame coefficients were not erased; with this information, the

signal can be estimated provided the corresponding frame elements span. In this sense, full spark frames are maximally robust to erasures. In particular, an $M \times N$ full spark frame is robust to $N - M$ erasures since any M of the frame coefficients will uniquely determine the original signal.

Yet another application of full spark frames is phaseless reconstruction, which can be viewed in terms of a channel, as in (4.1.2); in this case, $\mathcal{D}(\cdot)$ is the entrywise absolute value function. Phaseless reconstruction has a number of real-world applications including speech processing [BCE06], X-ray crystallography [CSV11], and quantum state estimation [RBKSC04]. As such, there has been a lot of work to reconstruct an M -dimensional vector (up to an overall phase factor) from the magnitudes of its frame coefficients, most of which involves frames in operator space, which inherently require $N = \Omega(M^2)$ measurements [BBCE09, RBKSC04]. However, Balan et al. [BCE06] show that if an $M \times N$ real frame F is full spark with $N \geq 2M - 1$, then $\mathcal{D} \circ F^*$ is injective, meaning an inversion process is possible with only $N = O(M)$ measurements. This result prompted an ongoing search for efficient phaseless reconstruction processes [BBCE07, CSV11], but no reconstruction process can succeed without a good family of frames, such as full spark frames.

Despite the fact that full spark frames have a multitude of applications, to date, there has not been much progress in constructing deterministic full spark frames, let alone full spark frames with additional desirable properties. A noteworthy exception is Püschel and Kovačević's work [PK05], in which real full spark tight frames are constructed using polynomial transforms. In the present paper, we start by investigating Vandermonde frames, harmonic frames, and modifications thereof. We find some classes of full spark ETFs, and characterize the $M \times N$ full spark harmonic frames in which N is a prime power. In Section 3, we show that $M \times N$ full spark Parseval frames form a dense subset of the entire collection of $M \times N$ Parseval frames. As such, full spark frames are abundant, even after imposing the additional condition

of tightness. This result is balanced with Section 4, in which we prove that verifying whether a matrix is full spark is hard for NP under randomized polynomial-time reductions. In other words, assuming $\text{NP} \not\subseteq \text{BPP}$ (a computational complexity assumption slightly stronger than $\text{P} \neq \text{NP}$ and nearly as widely believed), then there is no method by which one can efficiently test whether matrices are full spark. As such, the deterministic constructions in Section 2 are significant in that they guarantee a property which is otherwise difficult to check.

4.2 Deterministic constructions of full spark frames

A square matrix is invertible if and only if its determinant is nonzero, and in our quest for deterministic constructions of full spark frames, this characterization will reign supreme. One class of matrices has a particularly simple determinant formula: Vandermonde matrices. Specifically, Vandermonde matrices have the following form:

$$V = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \cdots & \alpha_N \\ \vdots & \vdots & \vdots & \vdots \\ \alpha_1^{M-1} & \alpha_2^{M-1} & \cdots & \alpha_N^{M-1} \end{bmatrix}, \quad (4.2.1)$$

and square Vandermonde matrices, i.e., with $N = M$, have the following determinant:

$$\det(V) = \prod_{1 \leq i < j \leq M} (\alpha_j - \alpha_i). \quad (4.2.2)$$

Consider (4.2.1) in the case where $N \geq M$. Since every $M \times M$ submatrix of V is also Vandermonde, we can modify the indices in (4.2.2) to calculate the determinant of the submatrices. These determinants are nonzero precisely when the bases $\{\alpha_n\}_{n=1}^N$ are distinct, yielding the following result:

Lemma 4.2.1. *A Vandermonde matrix is full spark if and only if its bases are distinct.*

When viewed in terms of frame theory, Vandermonde matrices naturally point to the discrete Fourier transform:

Theorem 4.2.2. *The only $M \times N$ Vandermonde matrices that are equal norm and tight have bases in the complex unit circle. Among these, the frames with the smallest worst-case coherence have bases that are equally spaced in the complex unit circle, provided $N \geq 2M$.*

Proof. Suppose a Vandermonde matrix is equal norm and tight. Note that a zero base will produce the zeroth identity basis element δ_0 . Letting $\mathcal{P}$ denote the indices of the nonzero bases, the fact that the matrix is full rank implies $|\mathcal{P}| \geq M - 1$. Also, equal norm gives that the frame element length

$$\|f_n\|^2 = \sum_{m=0}^{M-1} |f_n[m]|^2 = \sum_{m=0}^{M-1} |\alpha_n^m|^2 = \sum_{m=0}^{M-1} |\alpha_n|^{2m}$$

is constant over $n \in \mathcal{P}$. Since $\sum_{m=0}^{M-1} x^{2m}$ is strictly increasing over $0 < x < \infty$, there exists $c > 0$ such that $|\alpha_n|^2 = c$ for all $n \in \mathcal{P}$. Next, tightness gives that the rows have equal norm, implying that the first two rows have equal norm, i.e., $|\mathcal{P}|c = |\mathcal{P}|c^2$. Thus $c = 1$, and so the nonzero bases are in the complex unit circle. Furthermore, since the zeroth and first rows have equal norm by tightness, we have $|\mathcal{P}| = N$, and so every base is in the complex unit circle.

Now consider the inner product between Vandermonde frame elements whose bases $\{e^{2\pi i x_n}\}_{n=1}^N$ come from the complex unit circle:

$$\langle f_n, f_{n'} \rangle = \sum_{m=0}^{M-1} (e^{2\pi i x_n})^m \overline{(e^{2\pi i x_{n'}})^m} = \sum_{m=0}^{M-1} e^{2\pi i (x_n - x_{n'})m}.$$

We will show that the worst-case coherence comes from the two closest bases. Con-

sider the following function:

$$g(x) := \left| \sum_{m=0}^{M-1} e^{2\pi i x m} \right|^2.$$

We will prove two things about this function:

- (i) $\frac{d}{dx}g(x) < 0$ for every $x \in (0, \frac{1}{2M})$,
- (ii) $g(x) \leq g(\frac{1}{2M})$ for every $x \in (\frac{1}{2M}, 1 - \frac{1}{2M})$.

First, we claim that (i) and (ii) are sufficient to prove our result. To establish this, we first show that the two closest bases $e^{2\pi i x_{n'}}$ and $e^{2\pi i x_{n''}}$ satisfy $|x_{n'} - x_{n''}| \leq \frac{1}{2M}$. Without loss of generality, the n 's are ordered in such a way that $\{x_n\}_{n=0}^{N-1} \subseteq [0, 1)$ are nondecreasing. Define

$$d(x_n, x_{n+1}) := \begin{cases} x_{n+1} - x_n, & n = 0, \dots, N-2 \\ x_0 - (x_{N-1} - 1), & n = N-1, \end{cases}$$

and let n' be the n which minimizes $d(x_n, x_{n+1})$. Since the minimum is less than the average, we have

$$d(x_{n'}, x_{n'+1}) \leq \frac{1}{N} \left((x_0 - (x_{N-1} - 1)) + \sum_{n=0}^{N-1} (x_{n+1} - x_n) \right) = \frac{1}{N} \leq \frac{1}{2M}, \quad (4.2.3)$$

provided $N \geq 2M$. Note that if we view $\{x_n\}_{n \in \mathbb{Z}_N}$ as members of $\mathbb{R}/\mathbb{Z}$, then $d(x_n, x_{n+1}) = x_{n+1} - x_n$. Since $g(x)$ is even, then (i) implies that $|\langle f_{n'+1}, f_{n'} \rangle|^2 = g(x_{n'+1} - x_{n'})$ is larger than any other $g(x_p - x_{p'}) = |\langle f_p, f_{p'} \rangle|^2$ in which $x_p - x_{p'} \in [0, \frac{1}{2M}] \cup [1 - \frac{1}{2M}, 1)$. Next, (4.2.3) and (ii) together imply that $|\langle f_{n'+1}, f_{n'} \rangle|^2 = g(x_{n'+1} - x_{n'}) \geq g(\frac{1}{2M})$ is larger than any other $g(x_p - x_{p'}) = |\langle f_p, f_{p'} \rangle|^2$ in which $x_p - x_{p'} \in (\frac{1}{2M}, 1 - \frac{1}{2M})$, provided $N \geq 2M$. Combined, (i) and (ii) give that $|\langle f_{n'+1}, f_{n'} \rangle|$ achieves the worst-case coherence of $\{f_n\}_{n \in \mathbb{Z}_N}$. Additionally, (i) gives

that the worst-case coherence $|\langle f_{n'+1}, f_{n'} \rangle|$ is minimized when $x_{n'+1} - x_{n'}$ is maximized, i.e., when the x_n 's are equally spaced in the unit interval.

To prove (i), note that the geometric sum formula gives

$$g(x) = \left| \sum_{m=0}^{M-1} e^{2\pi i x m} \right|^2 = \left| \frac{e^{2M\pi i x} - 1}{e^{2\pi i x} - 1} \right|^2 = \frac{2 - 2\cos(2M\pi x)}{2 - 2\cos(2\pi x)} = \left(\frac{\sin(M\pi x)}{\sin(\pi x)} \right)^2, \quad (4.2.4)$$

where the final expression uses the identity $1 - \cos(2z) = 2\sin^2 z$. To show that g is decreasing over $(0, \frac{1}{2M})$, note that the base of (4.2.4) is positive on this interval, and performing the quotient rule to calculate its derivative will produce a fraction whose denominator is nonnegative and whose numerator is given by

$$M\pi \sin(\pi x) \cos(M\pi x) - \pi \sin(M\pi x) \cos(\pi x). \quad (4.2.5)$$

This factor is zero at $x = 0$ and has derivative:

$$-(M^2 - 1)\pi^2 \sin(\pi x) \sin(M\pi x),$$

which is strictly negative for all $x \in (0, \frac{1}{2M})$. Hence, (4.2.5) is strictly negative whenever $x \in (0, \frac{1}{2M})$, and so $g'(x) < 0$ for every $x \in (0, \frac{1}{2M})$.

For (ii), note that for every $x \in (\frac{1}{2M}, 1 - \frac{1}{2M})$, we can individually bound the numerator and denominator of what the geometric sum formula gives:

$$g(x) = \left| \sum_{m=0}^{M-1} e^{2\pi i x m} \right|^2 = \frac{|e^{2M\pi i x} - 1|^2}{|e^{2\pi i x} - 1|^2} \leq \frac{|e^{\pi i} - 1|^2}{|e^{\pi i/M} - 1|^2} = \left| \sum_{m=0}^{M-1} e^{\pi i m/M} \right|^2 = g(\frac{1}{2M}). \quad \square$$

Consider the $N \times N$ discrete Fourier transform (DFT) matrix, scaled to have

entries of unit modulus:

$$\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{N-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2(N-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \omega^{N-1} & \omega^{2(N-1)} & \cdots & \omega^{(N-1)(N-1)} \end{bmatrix},$$

where $\omega = e^{-2\pi i/N}$. The first M rows of the DFT form a Vandermonde matrix of distinct bases $\{\omega^n\}_{n=0}^{N-1}$; as such, this matrix is full spark by Lemma 4.2.1. In fact, the previous result says that this is in some sense an optimal Vandermonde frame, but this might not be the best way to pick rows from a DFT. Indeed, several choices of DFT rows could produce full spark frames, some with smaller coherence or other desirable properties, and so the remainder of this section focuses on full spark DFT submatrices. First, we note that not every DFT submatrix is full spark. For example, consider the 4×4 DFT:

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{bmatrix}.$$

Certainly, the zeroth and second rows of this matrix are not full spark, since the zeroth and second columns of this submatrix form the all-ones matrix, which is not invertible. So what can be said about the set of permissible row choices? The following result gives some necessary conditions on this set:

Theorem 4.2.3. *Take an $N \times N$ discrete Fourier transform matrix, and select the rows indexed by $\mathcal{M} \subseteq \mathbb{Z}_N$ to build the matrix F . If F is full spark, then so is the matrix built from rows indexed by*

- (i) *any translation of $\mathcal{M}$,*

(ii) any $A\mathcal{M}$ with A relatively prime to N ,

(iii) the complement of $\mathcal{M}$ in $\mathbb{Z}_N$.

Proof. For (i), we first define D to be the $N \times N$ diagonal matrix whose diagonal entries are $\{\omega^n\}_{n=0}^{N-1}$. Note that, since $\omega^{(m+1)n} = \omega^n \omega^{mn}$, translating the row indices $\mathcal{M}$ by 1 corresponds to multiplying F on the right by D . For some set $\mathcal{K} \subseteq \mathbb{Z}_N$ of size $M := |\mathcal{M}|$, let $F_{\mathcal{K}}$ denote the $M \times M$ submatrix of F whose columns are indexed by $\mathcal{K}$, and let $D_{\mathcal{K}}$ denote the $M \times M$ diagonal submatrix of D whose diagonal entries are indexed by $\mathcal{K}$. Then since $D_{\mathcal{K}}$ is unitary, we have

$$|\det((FD)_{\mathcal{K}})| = |\det(F_{\mathcal{K}}D_{\mathcal{K}})| = |\det(F_{\mathcal{K}})| |\det(D_{\mathcal{K}})| = |\det(F_{\mathcal{K}})|.$$

Thus, if F is full spark, $|\det((FD)_{\mathcal{K}})| = |\det(F_{\mathcal{K}})| > 0$, and so FD is also full spark. Using this fact inductively proves (i) for all translations of $\mathcal{M}$.

For (ii), let G denote the submatrix of rows indexed by $A\mathcal{M}$. Then for any set $\mathcal{K} \subseteq \mathbb{Z}_N$ of size M ,

$$\det(G_{\mathcal{K}}) = \det(\omega^{(Am)^k})_{m \in \mathcal{M}, k \in \mathcal{K}} = \det(\omega^{m(Ak)})_{m \in \mathcal{M}, k \in \mathcal{K}} = \det(F_{A\mathcal{K}}).$$

Since A is relatively prime to N , multiplication by A permutes the elements of $\mathbb{Z}_N$, and so $A\mathcal{K}$ has exactly M distinct elements. Thus, if F is full spark, then $\det(G_{\mathcal{K}}) = \det(F_{A\mathcal{K}}) \neq 0$, and so G is also full spark.

For (iii), we let G be the $(N - M) \times N$ submatrix of rows indexed by $\mathcal{M}^c$, so that

$$NI_N = \begin{bmatrix} F^* & G^* \end{bmatrix} \begin{bmatrix} F \\ G \end{bmatrix} = F^*F + G^*G. \quad (4.2.6)$$

We will use contraposition to show that F being full spark implies that G is also full spark. To this end, suppose G is not full spark. Then G has a collection of $N - M$

linearly dependent columns $\{g_i\}_{i \in \mathcal{K}}$, and so there exists a nontrivial sequence $\{\alpha_i\}_{i \in \mathcal{K}}$ such that

$$\sum_{i \in \mathcal{K}} \alpha_i g_i = 0.$$

Considering $g_i = G\delta_i$, where δ_i is the i th identity basis element, we can use (4.2.6) to express this linear dependence in terms of F :

$$0 = G^*0 = G^* \sum_{i \in \mathcal{K}} \alpha_i g_i = \sum_{i \in \mathcal{K}} \alpha_i G^* G \delta_i = \sum_{i \in \mathcal{K}} \alpha_i (N I_N - F^* F) \delta_i.$$

Rearranging then gives

$$x := N \sum_{i \in \mathcal{K}} \alpha_i \delta_i = \sum_{i \in \mathcal{K}} \alpha_i F^* F \delta_i. \quad (4.2.7)$$

Here, we note that x is nonzero since $\{\alpha_i\}_{i \in \mathcal{K}}$ is nontrivial, and that $x \in \text{Range}(F^* F)$.

Furthermore, whenever $j \notin \mathcal{K}$, we have from (4.2.7) that

$$\langle x, F^* F \delta_j \rangle = \langle F^* F x, \delta_j \rangle = N \left\langle F^* F \sum_{i \in \mathcal{K}} \alpha_i \delta_i, \delta_j \right\rangle = N^2 \left\langle \sum_{i \in \mathcal{K}} \alpha_i \delta_i, \delta_j \right\rangle = 0,$$

and so

$$x \perp \text{Span}\{F^* F \delta_j\}_{j \in \mathcal{K}^c}.$$

Thus, the containment

$$\text{Span}\{F^* F \delta_j\}_{j \in \mathcal{K}^c} \subseteq \text{Range}(F^* F)$$

is proper, and so

$$M = \text{Rank}(F) = \text{Rank}(F^* F) > \text{Rank}(F^* F_{\mathcal{K}^c}) = \text{Rank}(F_{\mathcal{K}^c}).$$

Since the $M \times M$ submatrix $F_{\mathcal{K}^c}$ is rank-deficient, it is not invertible, and therefore

F is not full spark. □

We note that our proof of (iii) above uses techniques from Cahill et al. [CCH11], and can be easily generalized to prove that the Naimark complement of a full spark tight frame is also full spark. Theorem 4.2.3 tells us quite a bit about the set of permissible choices for DFT rows. For example, not only can we pick the first M rows of the DFT to produce a full spark Vandermonde frame, but we can also pick any consecutive M rows, by Theorem 4.2.3(i). We would like to completely characterize the choices that produce full spark harmonic frames. The following classical result does this in the case where N is prime:

Theorem 4.2.4 (Chebotarëv, see [SL96]). *Let N be prime. Then every square submatrix of the $N \times N$ discrete Fourier transform matrix is invertible.*

As an immediate consequence of Chebotarëv's theorem, every choice of rows from the DFT produces a full spark harmonic frame, provided N is prime. Note that each of these are equal-norm tight frames by construction. Harmonic frames can also be designed to have minimal coherence; Xia et al. [XZG05] produces harmonic equiangular tight frames by selecting row indices which form a difference set in $\mathbb{Z}_N$. Interestingly, most known families of difference sets in $\mathbb{Z}_N$ require N to be prime [JPS07], and so the corresponding harmonic equiangular tight frames are guaranteed to be full spark by Chebotarëv's theorem. In the following, we use Chebotarëv's theorem to demonstrate full spark for a class of frames which contains harmonic frames, namely, frames which arise from concatenating harmonic frames with any number of identity basis elements:

Theorem 4.2.5 (cf. [Tao05, Theorem 1.1]). *Let N be prime, and pick any $M \leq N$ rows of the $N \times N$ discrete Fourier transform matrix to form the harmonic frame H . Next, pick any $K \leq M$, and take D to be the $M \times M$ diagonal matrix whose first K diagonal entries are $\sqrt{\frac{N+K-M}{MN}}$, and whose remaining $M - K$ entries are*

$\sqrt{\frac{N+K}{MN}}$. Then concatenating DH with the first K identity basis elements produces an $M \times (N + K)$ full spark unit norm tight frame.

Proof. Let F denote the resulting $M \times (N + K)$ frame. We start by verifying that F is unit norm. Certainly, the identity basis elements have unit norm. For the remaining frame elements, the modulus of each entry is determined by D , and so the norm squared of each frame element is

$$K\left(\frac{N+K-M}{MN}\right) + (M - K)\left(\frac{N+K}{MN}\right) = 1.$$

To demonstrate that F is tight, it suffices to show that $FF^* = \frac{N+K}{M}I_M$. The rows of DH are orthogonal since they are scaled rows of the DFT, while the rows of the identity portion are orthogonal because they have disjoint support. Thus, FF^* is diagonal. Moreover, the norm squared of each of the first K rows is $N\left(\frac{N+K-M}{MN}\right) + 1 = \frac{N+K}{M}$, while the norm squared of each of the remaining rows is $N\left(\frac{N+K}{MN}\right) = \frac{N+K}{M}$, and so $FF^* = \frac{N+K}{M}I_M$.

To demonstrate that F is full spark, first note that every $M \times M$ submatrix of DH is invertible since

$$|\det((DH)_\mathcal{K})| = |\det(DH_\mathcal{K})| = |\det(D)||\det(H_\mathcal{K})| > 0,$$

by Chebotarëv's theorem. Also, in the case where $K = M$, we note that the $M \times M$ submatrix of F composed solely of identity basis elements is trivially invertible. The only remaining case to check is when identity basis elements and columns of DH appear in the same $M \times M$ submatrix $F_\mathcal{K}$. In this case, we may shuffle the rows of $F_\mathcal{K}$ to have the form

$$\begin{bmatrix} A & 0 \\ B & I_K \end{bmatrix}.$$

Since shuffling rows has no impact on the size of the determinant, we may further use a determinant identity on block matrices to get

$$|\det(F_K)| = \left| \det \begin{bmatrix} A & 0 \\ B & I_K \end{bmatrix} \right| = |\det(A)\det(I_K)| = |\det(A)|.$$

Since A is a multiple of a square submatrix of the $N \times N$ DFT, we are done by Chebotarëv's theorem. $\square$

As an example of Theorem 4.2.5, pick N to be a prime congruent to 1 mod 4, and select $\frac{N+1}{2}$ rows of the $N \times N$ DFT according to the index set $\mathcal{M} := \{k^2 : k \in \mathbb{Z}_N\}$. If we take $K = 1$, the process in Theorem 4.2.5 produces an equiangular tight frame of redundancy 2, which can be verified using quadratic Gauss sums. Note that this corresponds to a special case of a construction in Zauner's thesis [Zau99], which was later studied by Renes [Ren07] and Strohmer [Str08]. Theorem 4.2.5 says that this construction is full spark.

Maximally sparse frames have recently become a subject of active research [CHKK11, FMT12]. We note that when $K = M$, Theorem 4.2.5 produces a maximally sparse $M \times (N + K)$ full spark frame, having a total of $M(M - 1)$ zero entries. To see that this sparsity level is maximal, we note that if the frame had any more zero entries, then at least one of the rows would have M zero entries, meaning the corresponding $M \times M$ submatrix would have a row of all zeros and hence a zero determinant. Similar ideas were studied previously by Nakamura and Masson [NM82].

Another interesting case is where $K = M = N$, i.e., when the frame constructed in Theorem 4.2.5 is a union of the unitary DFT and identity bases. Unions of orthonormal bases have received considerable attention in the context of sparse approximation [DE03, Tro08]. In fact, when N is a perfect square, concatenating the DFT with an identity basis forms the canonical example F of a dictionary with small spark [DE03]. To be clear, the Dirac comb of $\sqrt{N}$ spikes is an eigenvector of the

DFT, and so concatenating this comb with the negative of its Fourier transform produces a $2\sqrt{N}$ -sparse vector in the nullspace of F . In stark contrast, when N is prime, Theorem 4.2.5 shows that F is full spark.

The vast implications of Chebotarëv's theorem leads one to wonder whether the result admits any interesting generalization. Certainly, if one could characterize the full spark submatrices of a general DFT, this would provide ample freedom to optimize full spark frames for additional considerations. While we do not have a characterization for the general case, we do have one for the case where N is a prime power. Before stating the result, we require a definition:

Definition 4.2.6. We say a subset $\mathcal{M} \subseteq \mathbb{Z}_N$ is *uniformly distributed over the divisors of N* if, for every divisor d of N , the d cosets of $\langle d \rangle$ partition $\mathcal{M}$ into subsets, each of size $\lfloor \frac{|\mathcal{M}|}{d} \rfloor$ or $\lceil \frac{|\mathcal{M}|}{d} \rceil$.

At first glance, this definition may seem rather unnatural, but we will discover some important properties of uniformly distributed rows from the DFT. As an example, we take a short detour by considering the restricted isometry property (RIP), which has received considerable attention recently for its use in compressed sensing. We say a matrix F is (K, δ) -RIP if

$$(1 - \delta)\|x\|^2 \leq \|Fx\|^2 \leq (1 + \delta)\|x\|^2 \quad \text{whenever } \|x\|_0 \leq K.$$

Candès and Tao [CT05] demonstrated that the sparsest Ψ^*x which satisfies (4.1.3) can be found using ℓ_1 -minimization, provided F is $(2K, \sqrt{2} - 1)$ -RIP. Later, Rudelson and Vershynin [RV08] showed that a matrix of random rows from a DFT and normalized columns is RIP with high probability. We will show that harmonic frames satisfy RIP only if the selected row indices are nearly uniformly distributed over sufficiently small divisors of N .

To this end, recall that for any divisor d of N , the Fourier transform of the d -sparse

normalized Dirac comb $\frac{1}{\sqrt{d}}\chi_{\langle \frac{N}{d} \rangle}$ is the $\frac{N}{d}$ -sparse normalized Dirac comb $\sqrt{\frac{d}{N}}\chi_{\langle d \rangle}$. Let F be the $N \times N$ unitary DFT, and let H be the harmonic frame which arises from selecting rows of F indexed by $\mathcal{M}$ and then normalizing the columns. In order for H to be (K, δ) -RIP, $\mathcal{M}$ must contain at least one member of $\langle d \rangle$ for every divisor d of N which is $\leq K$, since otherwise

$$H \frac{1}{\sqrt{d}}\chi_{\langle \frac{N}{d} \rangle} = \sqrt{\frac{N}{|\mathcal{M}|}}(F \frac{1}{\sqrt{d}}\chi_{\langle \frac{N}{d} \rangle})_{\mathcal{M}} = \sqrt{\frac{N}{|\mathcal{M}|}}\left(\sqrt{\frac{d}{N}}\chi_{\langle d \rangle}\right)_{\mathcal{M}} = \sqrt{\frac{d}{|\mathcal{M}|}}\chi_{\mathcal{M} \cap \langle d \rangle} = 0,$$

which violates the lower RIP bound at $x = \frac{1}{\sqrt{d}}\chi_{\langle \frac{N}{d} \rangle}$. In fact, the RIP bounds indicate that

$$\|Hx\|^2 = \|H \frac{1}{\sqrt{d}}\chi_{\langle \frac{N}{d} \rangle}\|^2 = \left\| \sqrt{\frac{d}{|\mathcal{M}|}}\chi_{\mathcal{M} \cap \langle d \rangle} \right\|^2 = \frac{d}{|\mathcal{M}|}|\mathcal{M} \cap \langle d \rangle|$$

cannot be more than δ away from $\|x\|^2 = 1$. Similarly, taking x to be $\frac{1}{\sqrt{d}}\chi_{\langle \frac{N}{d} \rangle}$ modulated by a , i.e., $x[n] := \frac{1}{\sqrt{d}}\chi_{\langle \frac{N}{d} \rangle}[n]e^{2\pi i a n/N}$ for every $n \in \mathbb{Z}_N$, gives that $\|Hx\|^2 = \frac{d}{|\mathcal{M}|}|\mathcal{M} \cap (a + \langle d \rangle)|$ is also no more than δ away from 1. This observation gives the following result:

Theorem 4.2.7. *Select rows indexed by $\mathcal{M} \subseteq \mathbb{Z}_N$ from the $N \times N$ discrete Fourier transform matrix and then normalize the columns to produce the harmonic frame H . Then H satisfies the (K, δ) -restricted isometry property only if*

$$\left| |\mathcal{M} \cap (a + \langle d \rangle)| - \frac{|\mathcal{M}|}{d} \right| \leq \frac{|\mathcal{M}|}{d}\delta$$

for every divisor d of N with $d \leq K$ and every $a = 0, \dots, d-1$.

Now that we have an intuition for uniform distribution in terms of modulated Dirac combs and RIP, we take this condition to the extreme by considering uniform distribution over all divisors. Doing so produces a complete characterization of full spark harmonic frames when N is a prime power:

Theorem 4.2.8. *Let N be a prime power, and select rows indexed by $\mathcal{M} \subseteq \mathbb{Z}_N$ from the $N \times N$ discrete Fourier transform matrix to build the submatrix F . Then F is full spark if and only if $\mathcal{M}$ is uniformly distributed over the divisors of N .*

Note that, perhaps surprisingly, an index set $\mathcal{M}$ can be uniformly distributed over p but not over p^2 , and vice versa. For example, $\mathcal{M} = \{0, 1, 4\}$ is uniformly distributed over 2 but not 4, while $\mathcal{M} = \{0, 2\}$ is uniformly distributed over 4 but not 2.

Since the first M rows of a DFT form a full spark Vandermonde matrix, let's check that this index set is uniformly distributed over the divisors of N . For each divisor d of N , we partition the first M indices into the d cosets of $\langle d \rangle$. Write $M = qd + r$ with $0 \leq r < d$. The first qd of the M indices are distributed equally amongst all d cosets, and then the remaining r indices are distributed equally amongst the first r cosets. Overall, the first r cosets contain $q + 1 = \lfloor \frac{M}{d} \rfloor + 1$ indices, while the remaining $d - r$ cosets have $q = \lfloor \frac{M}{d} \rfloor$ indices; thus, the first M indices are indeed uniformly distributed over the divisors of N . Also, when N is prime, *every* subset of $\mathbb{Z}_N$ is uniformly distributed over the divisors of N in a trivial sense. In fact, Chebotarëv's theorem follows immediately from Theorem 4.2.8. In some ways, portions of our proof of Theorem 4.2.8 mirror recurring ideas in the existing proofs of Chebotarëv's theorem [DVB08, EI76, SL96, Tao05]. For the sake of completeness, we provide the full argument and save the reader from having to parse portions of proofs from multiple references. We start with the following lemmas, whose proofs are based on the proofs of Lemmas 1.2 and 1.3 in [Tao05].

Lemma 4.2.9. *Let N be a power of some prime p , and let $P(z_1, \dots, z_M)$ be a polynomial with integer coefficients. Suppose there exists N th roots of unity $\{\omega_m\}_{m=1}^M$ such that $P(\omega_1, \dots, \omega_M) = 0$. Then $P(1, \dots, 1)$ is a multiple of p .*

Proof. Denoting $\omega := e^{-2\pi i/N}$, then for every $m = 1, \dots, M$, we have $\omega_m = \omega^{k_m}$ for some $0 \leq k_m < N$. Defining the polynomial $Q(z) := P(z^{k_1}, \dots, z^{k_M})$, we have

$Q(\omega) = 0$ by assumption. Also, $Q(z)$ is a polynomial with integer coefficients, and so it must be divisible by the minimal polynomial of ω , namely, the cyclotomic polynomial $\Phi_N(z)$. Evaluating both polynomials at $z = 1$ then gives that $p = \Phi_N(1)$ divides $Q(1) = P(1, \dots, 1)$. $\square$

Lemma 4.2.10. *Let N be a power of some prime p , and pick $\mathcal{M} = \{m_i\}_{i=1}^M \subseteq \mathbb{Z}_N$ such that*

$$\frac{\prod_{1 \leq i < j \leq M} (m_j - m_i)}{\prod_{m=0}^{M-1} m!} \quad (4.2.8)$$

is not a multiple of p . Then the rows indexed by $\mathcal{M}$ in the $N \times N$ discrete Fourier transform form a full spark frame.

Proof. We wish to show that $\det(\omega_n^m)_{m \in \mathcal{M}, 1 \leq n \leq M} \neq 0$ for all M -tuples of distinct N th roots of unity $\{\omega_n\}_{n=1}^M$. Define the polynomial $D(z_1, \dots, z_M) := \det(z_n^m)_{m \in \mathcal{M}, 1 \leq n \leq M}$. Since columns i and j of $(z_n^m)_{m \in \mathcal{M}, 1 \leq n \leq M}$ are identical whenever $z_i = z_j$, we know that D vanishes in each of these instances, and so we can factor:

$$D(z_1, \dots, z_M) = P(z_1, \dots, z_M) \prod_{1 \leq i < j \leq M} (z_j - z_i)$$

for some polynomial $P(z_1, \dots, z_M)$ with integer coefficients. By Lemma 4.2.9, it suffices to show that $P(1, \dots, 1)$ is not a multiple of p , since this implies $D(\omega_1, \dots, \omega_M)$ is nonzero for all M -tuples of distinct N th roots of unity $\{\omega_n\}_{n=1}^M$.

To this end, we proceed by considering

$$A := \left(z_1 \frac{\partial}{\partial z_1} \right)^0 \left(z_2 \frac{\partial}{\partial z_2} \right)^1 \cdots \left(z_M \frac{\partial}{\partial z_M} \right)^{M-1} D(z_1, \dots, z_M) \Big|_{z_1 = \dots = z_M = 1}. \quad (4.2.9)$$

To compute A , we note that each application of $z_j \frac{\partial}{\partial z_j}$ produces terms according to the product rule. For some terms, a linear factor of the form $z_j - z_i$ or $z_i - z_j$ is replaced by z_j or $-z_j$, respectively. For each the other terms, these linear factors are untouched,

while another factor, such as $P(z_1, \dots, z_M)$, is differentiated and multiplied by z_j . Note that there are a total of $M(M-1)/2$ linear factors, and only $M(M-1)/2$ differentiation operators to apply. Thus, after expanding every product rule, there will be two types of terms: terms in which every differentiation operator was applied to a linear factor, and terms which have at least one linear factor remaining untouched. When we evaluate at $z_1 = \dots = z_M = 1$, the terms with linear factors vanish, and so the only terms which remain came from applying every differentiation operator to a linear factor. Furthermore, each of these terms before the evaluation is of the form $P(z_1, \dots, z_M) \prod_{1 \leq i < j \leq M} z_j$, and so evaluation at $z_1 = \dots = z_M = 1$ produces a sum of terms of the form $P(1, \dots, 1)$; to determine the value of A , it remains to count these terms. The $M-1$ copies of $z_M \frac{\partial}{\partial z_M}$ can only be applied to linear factors of the form $z_M - z_i$, of which there are $M-1$, and so there are a total of $(M-1)!$ ways to distribute these operators. Similarly, there are $(M-2)!$ ways to distribute the $M-2$ copies of $z_{M-1} \frac{\partial}{\partial z_{M-1}}$ amongst the $M-2$ linear factors of the form $z_{M-1} - z_i$. Continuing in this manner produces an expression for A :

$$A = (M-1)!(M-2)! \cdots 1!0! P(1, \dots, 1). \quad (4.2.10)$$

For an alternate expression of A , we substitute the definition of $D(z_1, \dots, z_M)$ into (4.2.9). Here, we exploit the multilinearity of the determinant and the fact that $(z_n \frac{\partial}{\partial z_n}) z_n^m = m z_n^m$ to get

$$A = \det(m^{n-1})_{m \in \mathcal{M}, 1 \leq n \leq M} = \prod_{1 \leq i < j \leq M} (m_j - m_i), \quad (4.2.11)$$

where the final equality uses the fact that $(m^{n-1})_{m \in \mathcal{M}, 1 \leq n \leq M}$ is the transpose of a Vandermonde matrix. Equating (4.2.10) to (4.2.11) reveals that (4.2.8) is an expression for $P(1, \dots, 1)$. Thus, by assumption, $P(1, \dots, 1)$ is not a multiple of p , and so we are done. $\square$

Proof of Theorem 4.2.8. ($\Leftarrow$) We will use Lemma 4.2.10 to demonstrate that F is full spark. To apply this lemma, we need to establish that (4.2.8) is not a multiple of p , and to do this, we will show that there are as many p -divisors in the numerator of (4.2.8) as there are in the denominator. We start by counting the p -divisors of the denominator:

$$\prod_{m=0}^{M-1} m! = \prod_{m=1}^{M-1} \prod_{\ell=1}^m \ell = \prod_{\ell=1}^{M-1} \prod_{m=1}^{M-\ell} \ell. \quad (4.2.12)$$

For each pair of integers $k, a \geq 1$, there are $\max\{M - ap^k, 0\}$ factors in (4.2.12) of the form $\ell = ap^k$. By adding these, we count each factor ℓ as many times as it can be expressed as a multiple of a power of p , which equals the number of p -divisors in ℓ . Thus, the number of p -divisors of (4.2.12) is

$$\sum_{k=1}^{\lfloor \log_p M \rfloor} \sum_{a=1}^{\lfloor \frac{M}{p^k} \rfloor} (M - ap^k). \quad (4.2.13)$$

Next, we count the p -divisors of the numerator of (4.2.8). To do this, we use the fact that $\mathcal{M}$ is uniformly distributed over the divisors of N . Since N is a power of p , the only divisors of N are smaller powers of p . Also, the cosets of $\langle p^k \rangle$ partition $\mathcal{M}$ into subsets $S_{k,b} := \{m_i \equiv b \pmod{p^k}\}$. We note that $m_j - m_i$ is a multiple of p^k precisely when m_i and m_j belong to the same subset $S_{k,b}$ for some $0 \leq b < p^k$. To count p -divisors, we again count each factor $m_j - m_i$ as many times as it can be expressed as a multiple of a prime power:

$$\sum_{k=1}^{\lfloor \log_p M \rfloor} \sum_{b=0}^{p^k-1} \binom{|S_{k,b}|}{2}. \quad (4.2.14)$$

Write $M = qp^k + r$ with $0 \leq r < p^k$. Then $q = \lfloor \frac{M}{p^k} \rfloor$. Since $\mathcal{M}$ is uniformly distributed over p^k , there are r subsets $S_{k,b}$ with $q+1$ elements and $p^k - r$ subsets with q elements.

We use this to get

$$\sum_{b=0}^{p^k-1} \binom{|S_{k,b}|}{2} = \binom{q+1}{2} r + \binom{q}{2} (p^k - r) = \frac{q}{2} ((q-1)p^k + 2r + (qp^k - qp^k)).$$

Rearranging and substituting $M = qp^k + r$ then gives

$$\sum_{b=0}^{p^k-1} \binom{|S_{k,b}|}{2} = \frac{q}{2} (2M - (q+1)p^k) = Mq - \binom{q+1}{2} p^k = \sum_{a=1}^{\lfloor \frac{M}{p^k} \rfloor} (M - ap^k).$$

Thus, there are as many p -divisors in the numerator (4.2.14) as there are in the denominator (4.2.13), and so (4.2.8) is not divisible by p . Lemma 4.2.10 therefore gives that F is full spark.

($\Rightarrow$) We will prove that this direction holds regardless of whether N is a prime power. Suppose $\mathcal{M} \subseteq \mathbb{Z}_N$ is not uniformly distributed over the divisors of N . Then there exists a divisor d of N such that one of the cosets of $\langle d \rangle$ intersects $\mathcal{M}$ with $\leq \lfloor \frac{M}{d} \rfloor - 1$ or $\geq \lceil \frac{M}{d} \rceil + 1$ indices. Notice that if a coset of $\langle d \rangle$ intersects $\mathcal{M}$ with $\leq \lfloor \frac{M}{d} \rfloor - 1$ indices, then the complement $\mathcal{M}^c$ intersects the same coset with $\geq \lceil \frac{N-M}{d} \rceil + 1 = \lceil \frac{|\mathcal{M}^c|}{d} \rceil + 1$ indices. By Theorem 4.2.3(iii), $\mathcal{M}$ produces a full spark harmonic frame precisely when $\mathcal{M}^c$ produces a full spark harmonic frame, and so we may assume without loss of generality that there exists a coset of $\langle d \rangle$ which intersects $\mathcal{M}$ with $\geq \lceil \frac{M}{d} \rceil + 1$ indices.

To prove that the rows with indices in $\mathcal{M}$ are not full spark, we find column entries which produce a singular submatrix. Writing $M = qd + r$ with $0 \leq r < d$, let $\mathcal{K}$ contain $q = \lfloor \frac{M}{d} \rfloor$ cosets of $\langle \frac{N}{d} \rangle$ along with r elements from an additional coset. We claim that the DFT submatrix with row entries $\mathcal{M}$ and column entries $\mathcal{K}$ is singular. To see this, shuffle the rows and columns to form a matrix A in which the row entries are grouped into common cosets of $\langle d \rangle$ and the column entries are grouped into common cosets of $\langle \frac{N}{d} \rangle$. This breaks A into rank-1 submatrices: each

pair of cosets $a + \langle d \rangle$ and $b + \langle \frac{N}{d} \rangle$ produces a submatrix

$$(\omega^{(a+id)(b+j\frac{N}{d})})_{i \in \mathcal{I}, j \in \mathcal{J}} = \omega^{ab} (\omega^{bdi} \omega^{a\frac{N}{d}j})_{i \in \mathcal{I}, j \in \mathcal{J}}$$

for some index sets $\mathcal{I}$ and $\mathcal{J}$; this is a rank-1 outer product. Let $\mathcal{L}$ be the largest intersection between $\mathcal{M}$ and a coset of $\langle d \rangle$. Then $|\mathcal{L}| \geq \lceil \frac{M}{d} \rceil + 1$ is the number of rows in the tallest of these rank-1 submatrices. Define $A_{\mathcal{L}}$ to be the $M \times M$ matrix with entries $A_{\mathcal{L}}[i, j] = A[i, j]$ whenever $i \in \mathcal{L}$ and zero otherwise. Then

$$\text{Rank}(A) = \text{Rank}(A_{\mathcal{L}} + A - A_{\mathcal{L}}) \leq \text{Rank}(A_{\mathcal{L}}) + \text{Rank}(A - A_{\mathcal{L}}). \quad (4.2.15)$$

Since $A - A_{\mathcal{L}}$ has $|\mathcal{L}|$ rows of zero entries, we also have

$$\text{Rank}(A - A_{\mathcal{L}}) \leq M - |\mathcal{L}| \leq M - (\lceil \frac{M}{d} \rceil + 1). \quad (4.2.16)$$

Moreover, since we can decompose $A_{\mathcal{L}}$ into a sum of $\lceil \frac{M}{d} \rceil$ zero-padded rank-1 submatrices, we have $\text{Rank}(A_{\mathcal{L}}) \leq \lceil \frac{M}{d} \rceil$. Combining this with (4.2.15) and (4.2.16) then gives that $\text{Rank}(A) \leq M - 1$, and so the DFT submatrix is not invertible. $\square$

Note that our proof of Theorem 4.2.8 establishes the necessity of having row indices uniformly distributed over the divisors of N in the general case. This leaves some hope for completely characterizing full spark harmonic frames. Naturally, one might suspect that the uniform distribution condition is sufficient in general, but this suspicion fails when $N = 10$. Indeed, the following DFT submatrix is singular despite the row indices being uniformly distributed over the divisors of 10:

$$(e^{-2\pi i mn/10})_{m \in \{0,1,3,4\}, n \in \{0,1,2,6\}}.$$

Also, just as we used Chebotarëv's theorem to analyze the harmonic equiangular

tight frames from Xia et al. [XZG05], we can also use Theorem 4.2.8 to determine whether harmonic equiangular tight frames with a prime power number of frame elements are full spark. Unfortunately, none of the infinite families in [XZG05] have the number of frame elements in the form of a prime power (other than primes). Luckily, there is at least one instance in which the number of frame elements happens to be a prime power: the harmonic frames that arise from Singer difference sets have $M = \frac{q^d-1}{q-1}$ and $N = \frac{q^{d+1}-1}{q-1}$ for a prime power q and an integer $d \geq 2$; when $q = 3$ and $d = 4$, the number of frame elements $N = 11^2$ is a prime power. In this case, the row indices we select are

$$\mathcal{M} = \{1, 2, 3, 6, 7, 9, 11, 18, 20, 21, 25, 27, 33, 34, 38, 41, 44, 47, 53, 54, 55, 56, \\ 58, 59, 60, 63, 64, 68, 70, 71, 75, 81, 83, 89, 92, 99, 100, 102, 104, 114\},$$

but these are not uniformly distributed over 11, and so the corresponding harmonic frame is not full spark by Theorem 4.2.8.

4.3 Full spark frames are dense

In this section, we rigorously show that full spark frames are dense in the entire set of matrices. This corresponds to our intuition that a matrix whose entries are independent continuous random variables is full spark with probability one. Later, we apply the additional constraint of tightness, specifically Parseval tightness, where the frame bound is 1. With this constraint, we will show that full spark Parseval frames are dense in the entire set of Parseval frames.

In order to make our arguments rigorous, we view each entry F_{mn} of the $M \times N$ matrix F in terms of its real and imaginary parts: $x_{mn} + iy_{mn}$; this decomposition will become helpful later when we consider inner products between rows of F , which are not algebraic operations on complex vectors. Recall that F is full spark pre-

cisely when each $M \times M$ submatrix has nonzero determinant. We note that for each submatrix, the determinant is a polynomial in the x_{mn} 's and y_{mn} 's, and having this polynomial be nonzero is equivalent to having either its real or imaginary part be nonzero. This naturally leads us to the following definition from algebraic geometry: A *real algebraic variety* is the set of common zeros of a finite set of polynomials, that is, given polynomials $p_1, \dots, p_r \in \mathbb{R}[x_1, \dots, x_k]$, we define the corresponding real algebraic variety by

$$V(p_1, \dots, p_r) := \{x \in \mathbb{R}^k : p_1(x) = \dots = p_r(x) = 0\}.$$

Each submatrix determinant corresponds to a real algebraic variety $V \subseteq \mathbb{R}^{2MN}$ of two polynomials, and having this determinant be nonzero is equivalent to restricting to the complement of V .

In general, a variety is equipped with a topology known as the *Zariski topology*, in which subvarieties are the closed sets. As such, the set of matrices with a nonzero determinant forms a Zariski-open set, since it is the complement of a variety. In order to exploit this Zariski-openness, we require the additional concept of irreducibility: A variety is said to be *irreducible* if it cannot be written as a finite union of proper subvarieties. As an example, the entire space $\mathbb{R}^k$ is the variety which corresponds to the zero polynomial; in this case, every proper subvariety is lower-dimensional, and so $\mathbb{R}^k$ is trivially irreducible. On the other hand, the variety in $\mathbb{R}^2$ defined by $xy = 0$ is not irreducible because it can be expressed as the union of varieties defined by $x = 0$ and $y = 0$. Irreducibility is important because it says something about Zariski-open sets:

Theorem 4.3.1. *If V is an irreducible algebraic variety, then every nonempty Zariski-open subset of V is dense in V in the standard topology.*

For example, the hyperplane defined by $x_1 = 0$ is a subvariety of $\mathbb{R}^k$, and since $\mathbb{R}^k$

is irreducible, the complement of the hyperplane is dense in the standard topology. Going back to the variety defined by $xy = 0$, we know it can be expressed as a union of proper subvarieties, namely the x - and y -axes, and complementing one gives a subset of the other, neither of which is dense in the entire variety.

We are now ready to prove the following result:

Theorem 4.3.2. *Every matrix is arbitrarily close to a full spark frame.*

Proof. In an $M \times N$ matrix, there are $\binom{N}{M}$ submatrices of size $M \times M$, and the determinant of each of these submatrices corresponds to a variety of two real polynomials. Since the set of $M \times N$ full spark frames is defined as the (finite) intersection of the complements of these varieties, it is a Zariski-open subset of the irreducible variety $\mathbb{R}^{2MN}$ of all $M \times N$ matrices. Moreover, this set is nonempty since it contains the matrix formed by the first M rows of the $N \times N$ DFT, and so we are done by Theorem 4.3.1. $\square$

We now focus on the set of *Parseval frames*, that is, tight frames with frame bound 1. Note that $M \times N$ Parseval frames are characterized by the rows forming an orthonormal system of size M in N -dimensional space. The set of all such orthonormal systems is known as the *Stiefel manifold*, denoted $\text{St}(M, N)$. In general, a *manifold* is a set of vectors with a well-defined tangent space at every point in the set. Since this manifold property is nice, we would like to think of varieties as manifolds, but there exist varieties with points at which tangent spaces are not well-defined; such points are called *singularities*. Note that our definition of singularity is geometric, i.e., where the variety fails to be a real differentiable manifold, as opposed to algebraic. For example, the variety defined by $xy = 0$ is the union of the x - and y -axes, and as such, has a singularity at $x = y = 0$. Certainly with different types of varieties, there are other types of singularities which may arise, but there are also many varieties which do not have singularities at all, and we call these varieties *nonsingular*. Nonsingularity is a

useful property for the following reason:

Theorem 4.3.3. *An algebraic variety which is nonsingular and connected is necessarily irreducible.*

Note that the orthonormality conditions which characterize Parseval frames F can be expressed as polynomial equations in the real and imaginary parts of the entries F_{mn} . As such, we may view the Stiefel manifold as a real algebraic variety; this variety is nonsingular because it is a manifold. Moreover, the variety is connected because the (connected) unitary group $U(N)$ acts transitively on $\text{St}(M, N)$. Thus by Theorem 4.3.3, the variety of Parseval frames is irreducible. Having established this, we can now prove the following result:

Theorem 4.3.4. *Every Parseval frame is arbitrarily close to a full spark Parseval frame.*

Proof. Proceeding as in the proof of Theorem 4.3.2, we have that the set of $M \times N$ full spark Parseval frames is Zariski-open in the irreducible variety of all $M \times N$ Parseval frames. Again considering the first M rows of the $N \times N$ DFT, we know this set is nonempty, and so we are done by Theorem 4.3.1. $\square$

We note that Theorems 4.3.2 and 4.3.4 are also true when we further require the frames to be real. In this case, we cannot use the first M rows of the $N \times N$ DFT to establish that the Zariski-open sets are nonempty. For the real version of Theorem 4.3.2, we use an $M \times N$ Vandermonde matrix with distinct real bases; see Lemma 4.2.1. However, this construction must be modified to use it in the proof of the real version of Theorem 4.3.4, since such Vandermonde matrices will not be tight; see Theorem 4.2.2. Given an $M \times N$ full spark frame F , the modification $G := (FF^*)^{-1/2}F$ is full spark and Parseval; indeed, $GG^* = I_M$, and since F is a frame, $(FF^*)^{-1/2}$ is full rank, and so the columns of G are linearly independent precisely when the corresponding columns of F are linearly independent.

Another way that the proof of Theorem 4.3.4 changes in the real case is in verifying that the real Stiefel manifold is irreducible. Just as in the complex case, this follows from Theorem 4.3.3 since $\text{St}(M, N)$ is connected [Rap02], but the fact that $\text{St}(M, N)$ is connected in the real case is not immediate. By analogy, the orthogonal group $O(N)$ certainly acts transitively on $\text{St}(M, N)$, but unlike the unitary group, the orthogonal group has two connected components. Intuitively, this is resolved by the fact that $N > M$, granting additional freedom of movement throughout $\text{St}(M, N)$.

In addition to Theorems 4.3.2 and 4.3.4, we would like a similar result for unit norm tight frames, i.e., that every unit norm tight frame is arbitrarily close to a full spark unit norm tight frame. Certainly, the set of unit norm tight frames is a real algebraic variety, but it is unclear whether this variety is irreducible. Without knowing whether the variety is irreducible, we can follow the proofs of Theorems 4.3.2 and 4.3.4 to conclude that full spark unit norm tight frames are dense in the irreducible components in which they exist—a far cry from the density result we seek. This illustrates a significant gap in our current understanding of the variety of unit norm tight frames. It should be mentioned that Strawn [Str11] showed that the variety of $M \times N$ unit norm tight frames (over real or complex space) is nonsingular precisely when M and N are relatively prime. Additionally, Dykema and Strawn [DS06] proved that the variety of $2 \times N$ real unit norm tight frames is connected, and so by Theorem 4.3.3, this variety is irreducible when N is odd. It is unknown whether the variety of $M \times N$ unit norm tight frames is connected in general. Finally, we note that a Theorem 4.3.4 gives a weaker version of the result we would like: every unit norm tight frame is arbitrarily close to a full spark tight frame with frame element lengths arbitrarily close to 1.

4.4 The computational complexity of verifying full spark

In the previous section, we demonstrated the abundance of full spark frames, even after imposing the additional condition of tightness. But how would we check whether a particular $M \times N$ frame is full spark? Naturally, we would determine whether each of the $M \times M$ submatrices is invertible; computing $\binom{N}{M}$ determinants would do. However, this is not necessarily the fastest way to go about testing whether a matrix is full spark. For example, perhaps there is an easy-to-calculate expression for the product of the $\binom{N}{M}$ determinants; after all, this product is nonzero precisely when the matrix is full spark. If such an expression exists, then it would suffice to find it, but how might one disprove the existence of any such expression?

At the heart of the matter is computational complexity theory, which provides a rigorous playing field for expressing how hard certain problems are. In this section, we consider the complexity of the following problem:

Problem 4.4.1 (FULL SPARK). Given a matrix, is it full spark?

As an example of complexity, intuitively, doubling an integer is no harder than adding integers, since one can use addition to multiply by 2. To make this more precise, complexity theorists use what is called a *polynomial-time reduction*, that is, a polynomial-time algorithm that solves problem A by exploiting an oracle which solves problem B ; the reduction indicates that solving problem A is no harder than solving problem B (up to polynomial factors in time), and we say “ A reduces to B ,” or $A \leq B$. Since we can use the polynomial-time routine $x + x$ to produce $2x$, we conclude that doubling an integer reduces to adding integers, as expected.

In complexity theory, problems are categorized into complexity classes according to the amount of resources required to solve them. For example, the complexity class P contains all problems which can be solved in polynomial time, while problems

in EXPTIME may require as much as exponential time. Problems in NP have the defining quality that solutions can be verified in polynomial time given a certificate for the answer. Clearly, $P \subseteq NP$, since we can ignore the certificate and still solve the problem in polynomial time. Finally, a problem B is called NP-*hard* if every problem A in NP reduces to B , that is, every problem in NP is no harder than B , and a problem is called NP-*complete* if it is both NP-hard and in NP. We note that $A \leq B$ and $B \leq C$ together imply $A \leq C$, and so to demonstrate that a problem C is NP-hard, it suffices to show that $B \leq C$ for some NP-hard problem B . We also note that an NP-hard problem can be solved in polynomial time only if $P = NP$; this is an open problem, but it is widely believed that $P \neq NP$.

Sometimes, it can be difficult to find a deterministic reduction from one problem to another. One example is reducing the satisfiability problem (SAT) to the unique satisfiability problem (UNIQUE SAT). To be clear, SAT is an NP-hard problem [Kar72] that asks whether there exists an input for which a given Boolean function returns “true,” while UNIQUE SAT asks the same question with an additional promise: that the given Boolean function is satisfiable only if there is a *unique* input for which it returns “true.” Intuitively, UNIQUE SAT is easier than SAT because we might be able to exploit the additional structure of uniquely satisfiable Boolean functions; thus, it could be difficult to find a reduction from SAT to UNIQUE SAT. Despite this intuition, there is a *randomized* polynomial-time reduction from SAT to UNIQUE SAT [VV86]. Defined over all Boolean functions of n variables, the reduction maps functions that are not satisfiable to other functions that are not satisfiable, and with probability $\geq \frac{1}{8n}$, it maps satisfiable functions to uniquely satisfiable functions. After applying this reduction to a given Boolean function, if a UNIQUE SAT oracle declares “uniquely satisfiable,” then we know for certain that the original Boolean function was satisfiable. But the reduction will only map a satisfiable problem to a uniquely satisfiable problem with probability $\geq \frac{1}{8n}$, so what good is this reduction?

The answer lies in something called *amplification*; since the success probability is, at worst, polynomially small in n (i.e., $\geq \frac{1}{p(n)}$), we can repeat our oracle-based randomized algorithm a polynomial number of times $np(n)$ and achieve an error probability $\leq (1 - \frac{1}{p(n)})^{np(n)} \sim e^{-n}$ which is exponentially small. In this section, we give a randomized polynomial-time reduction from the following problem:

Problem 4.4.2. Given a bipartite graph, what is the girth of its transversal matroid?

Before giving the reduction, we will show that Problem 4.4.2 is NP-hard, but first we briefly review some matroid theory. A *matroid* is a pair $(E, \mathcal{I})$, where E is a finite set and $\mathcal{I}$ is a collection of *independent* subsets of E such that

- (i) $\emptyset \in \mathcal{I}$,
- (ii) $A \subseteq B \in \mathcal{I}$ implies $A \in \mathcal{I}$,
- (iii) $A, B \in \mathcal{I}$ and $|A| > |B|$ implies $\exists x \in A \setminus B$ such that $B \cup \{x\} \in \mathcal{I}$.

We see that matroids provide a natural generalization of the concept of linear independence in vector spaces. Just as spark is the size of the smallest linearly dependent set, the *girth* of a matroid is the size of the smallest subset of E that is not in $\mathcal{I}$. For different kinds of matroids, there are different ways to identify the independent sets. For a *transversal* matroid, the natural way to define $\mathcal{I}$ uses a bipartite graph between two disjoint sets of vertices in which the vertices in one of the disjoint sets are labeled by elements of E ; in this case, $\mathcal{I}$ is defined to be all subsets of E whose vertices form the ends of a matching. Hall's marriage theorem [Hal35] gives a remarkable characterization of the independent sets in a transversal matroid: $B \in \mathcal{I}$ if and only if every subset $A \subseteq B$ has $\geq |A|$ neighbors in the bipartite graph. A matroid $(E, \mathcal{I})$ is *representable over a field* $\mathbb{F}$ if, for some M , there exists a mapping $\varphi: E \rightarrow \mathbb{F}^M$ such that $\varphi(A)$ is linearly independent if and only if $A \in \mathcal{I}$; as such, the girth of $(E, \mathcal{I})$ is the spark of $\varphi(E)$. In our reduction, we make use of the fact that every transversal matroid is representable over $\mathbb{R}$ [PW70].

The following result comes from McCormick's thesis [McC83], which credits the proof to Stockmeyer; since [McC83] is difficult to access and the proof is instructive, we include it below:

Theorem 4.4.3. *Problem 4.4.2 is NP-hard.*

Proof. We will reduce from the NP-complete clique decision problem, which asks “Given a graph, does it contain a clique of K vertices?” [Kar72]. First, we may assume $K \geq 4$ without loss of generality, since any such clique can otherwise be found in cubic time by an exhaustive search. Take a graph $G = (E, V)$, and consider the bipartite graph G' between disjoint sets E and $V \sqcup \{1, \dots, \binom{K}{2} - K - 1\}$, in which $e \leftrightarrow v$ for every $e \in E$ and $v \in e$, and $e \leftrightarrow k$ for every $e \in E$ and $k \in \{1, \dots, \binom{K}{2} - K - 1\}$. We claim that the girth of the transversal matroid of G' is $\binom{K}{2}$ precisely when there exists a K -clique in G .

We start by analyzing the girth of a transversal matroid. Consider any dependent set $C \subseteq E$ with $\leq |C| - 2$ neighbors in G' . Then removing any member x of C will produce a smaller set $C \setminus \{x\}$ with $\leq |C| - 2 = |C \setminus \{x\}| - 1$ neighbors in G' , which is necessarily dependent by the pigeonhole principle. Now consider any dependent set $C \subseteq E$ with $\geq |C|$ neighbors in G' . By the Hall's marriage theorem, there exists a proper subset $C' \subseteq C$ with $< |C'|$ neighbors in G' , meaning C' is a smaller dependent set. Thus, the girth c is the size of the smallest subset $C \subseteq E$ with $|C| - 1$ total neighbors in G' .

Suppose $c = \binom{K}{2}$. Then since C is adjacent to every vertex in $\{1, \dots, \binom{K}{2} - K - 1\}$, C has $(c - 1) - (\binom{K}{2} - K - 1) = K$ neighbors in V . These are precisely the vertices D in G which are induced by the edges in C , and so C is contained in the set C' of edges induced by D , of which there are $\leq \binom{|D|}{2} = \binom{K}{2}$, with equality only if D induces a K -clique in G . Since $\binom{K}{2} = |C| \leq |C'| \leq \binom{|D|}{2} = \binom{K}{2}$ implies equality, there exists a K -clique in G .

Now suppose there exists a K -clique with edges C . Then C has $\binom{K}{2}$ elements and

$K + \left(\binom{K}{2} - K - 1\right) = \binom{K}{2} - 1$ neighbors in G' . To prove that $c = \binom{K}{2}$, it suffices to show that there is no smaller subset $C' \subseteq E$ with $|C'| - 1$ total neighbors in G' . Suppose, to the contrary, that $c = \binom{K}{2} - \ell$ for some $\ell > 0$. Then there exists $C' \subseteq E$ with c elements and $(c - 1) - \left(\binom{K}{2} - K - 1\right) = K - \ell$ neighbors in V . Note that each $e \in E$ contains two vertices in G , and so by the definition of G' , C' necessarily has $K - \ell \geq 2$ neighbors in V . Also, since the $K - \ell$ neighbors of C' in V arise from the subgraph of G induced by C' , and since those neighbors induce at most $\binom{K-\ell}{2}$ edges including C' , we have $\binom{K}{2} - \ell = c \leq \binom{K-\ell}{2}$. This inequality simplifies to $\ell \geq 2K - 3$, which combines with $K - \ell \geq 2$ to contradict the fact that $\ell > 0$. $\square$

Having established that Problem 4.4.2 is NP-hard, we reduce from it the main problem of this section. Our proof is specifically geared toward the case where the matrix in question has integer entries; this is stronger than manipulating real (complex) numbers exactly as well as with truncations and tolerances.

Theorem 4.4.4. *FULL SPARK is hard for NP under randomized polynomial-time reductions.*

Proof. We will give a randomized polynomial-time reduction from Problem 4.4.2 to FULL SPARK. As such, suppose we are given a bipartite graph G , in which every edge is between the disjoint sets A and B . Take $M := |B|$ and $N := |A|$. Using this graph, we randomly draw an $M \times N$ matrix F using the following process: for each $i \in B$ and $j \in A$, pick the entry F_{ij} randomly from $\{1, \dots, N2^{N+1}\}$ if $i \leftrightarrow j$ in G ; otherwise set $F_{ij} = 0$. In Proposition 3.11 of [Mar09], it is shown that the columns of F form a representation of the transversal matroid of G with probability $\geq \frac{1}{2}$. For the moment, we assume that F succeeds in representing the matroid.

Since the girth of the original matroid equals the spark of its representation, for each $K = 1, \dots, M$, we test whether $\text{Spark}(F) > K$. To do this, take H to be some $M \times P$ full spark frame. We will determine an appropriate value for P

later, but for simplicity, we can take H to be the Vandermonde matrix formed from bases $\{1, \dots, P\}$; see Lemma 4.2.1. We claim we can randomly select K indices $\mathcal{K} \subseteq \{1, \dots, P\}$ and test whether $H_{\mathcal{K}}^* F$ is full spark to determine whether $\text{Spark}(F) > K$. Moreover, after performing this test for each $K = 1, \dots, M$, the probability of incorrectly determining $\text{Spark}(F)$ is $\leq \frac{1}{2}$, provided P is sufficiently large.

We want to test whether $H_{\mathcal{K}}^* F$ is full spark and use the result as a proxy for whether $\text{Spark}(F) > K$. For this to work, we need to have $\text{Rank}(H_{\mathcal{K}}^* F_{\mathcal{K}'}) = K$ precisely when $\text{Rank}(F_{\mathcal{K}'}) = K$ for every $\mathcal{K}' \subseteq \{1, \dots, N\}$ of size K . To this end, it suffices to have the nullspace $\mathcal{N}(H_{\mathcal{K}}^*)$ of $H_{\mathcal{K}}^*$ intersect trivially with the column space of $F_{\mathcal{K}'}$ for every $\mathcal{K}'$. To be clear, it is always the case that $\text{Rank}(H_{\mathcal{K}}^* F_{\mathcal{K}'}) \leq \text{Rank}(F_{\mathcal{K}'})$, and so $\text{Rank}(F_{\mathcal{K}'}) < K$ implies $\text{Rank}(H_{\mathcal{K}}^* F_{\mathcal{K}'}) < K$. If we further assume that $\mathcal{N}(H_{\mathcal{K}}^*) \cap \text{Span}(F_{\mathcal{K}'}) = \{0\}$, then the converse also holds. To see this, suppose $\text{Rank}(H_{\mathcal{K}}^* F_{\mathcal{K}'}) < K$. Then by the rank-nullity theorem, there is a nontrivial $x \in \mathcal{N}(H_{\mathcal{K}}^* F_{\mathcal{K}'})$. Since $H_{\mathcal{K}}^* F_{\mathcal{K}'} x = 0$, we must have $F_{\mathcal{K}'} x \in \mathcal{N}(H_{\mathcal{K}}^*)$, which in turn implies $x \in \mathcal{N}(F_{\mathcal{K}'})$ since $\mathcal{N}(H_{\mathcal{K}}^*) \cap \text{Span}(F_{\mathcal{K}'}) = \{0\}$ by assumption. Thus, $\text{Rank}(F_{\mathcal{K}'}) < K$ by the rank-nullity theorem.

Now fix $\mathcal{K}' \subseteq \{1, \dots, N\}$ of size K such that $\text{Rank}(F_{\mathcal{K}'}) = K$. We will show that the vast majority of choices $\mathcal{K} \subseteq \{1, \dots, P\}$ of size K satisfy $\mathcal{N}(H_{\mathcal{K}}^*) \cap \text{Span}(F_{\mathcal{K}'}) = \{0\}$. To do this, we consider the columns $\{h_k\}_{k \in \mathcal{K}}$ of $H_{\mathcal{K}}$ one at a time, and we make use of the fact that $\mathcal{N}(H_{\mathcal{K}}^*) = \bigcap_{k \in \mathcal{K}} \mathcal{N}(h_k^*)$. In particular, since H is full spark, there are at most $M - K$ columns of H in the orthogonal complement of $\text{Span}(F_{\mathcal{K}'})$, and so there are at least $P - (M - K)$ choices of h_{k_1} for which $\mathcal{N}(h_{k_1}^*)$ does not contain $\text{Span}(F_{\mathcal{K}'})$, i.e.,

$$\dim(\mathcal{N}(h_{k_1}^*) \cap \text{Span}(F_{\mathcal{K}'})) = K - 1.$$

Similarly, after selecting the first J h_k 's, we have $\dim(S) = K - J$, where

$$S := \bigcap_{j=1}^J \mathcal{N}(h_{k_j}^*) \cap \text{Span}(F_{\mathcal{K}'})$$

Again, since H is full spark, there are at most $M - (K - J)$ columns of H in the orthogonal complement of S , and so the remaining $P - (M - (K - J))$ columns are candidates for $h_{k_{J+1}}$ that give

$$\dim\left(\bigcap_{j=1}^{J+1} \mathcal{N}(h_{k_j}^*) \cap \text{Span}(F_{\mathcal{K}'})\right) = \dim\left(\mathcal{N}(h_{k_{J+1}}^*) \cap S\right) = K - (J + 1).$$

Overall, if we randomly pick $\mathcal{K} \subseteq \{1, \dots, P\}$ of size K , then

$$\begin{aligned} \Pr\left(\mathcal{N}(H_{\mathcal{K}}^*) \cap \text{Span}(F_{\mathcal{K}'}) = \{0\}\right) &\geq \left(1 - \frac{M-K}{P}\right)\left(1 - \frac{M-(K-1)}{P}\right) \cdots \left(1 - \frac{M-1}{P}\right) \\ &\geq \left(1 - \frac{M}{P}\right)^K \\ &\geq 1 - \frac{MK}{P}, \end{aligned}$$

where the final step is by Bernoulli's inequality. Taking a union bound over all choices of $\mathcal{K}' \subseteq \{1, \dots, N\}$ and all values of $K = 1, \dots, M$ then gives

$$\begin{aligned} \Pr\left(\begin{array}{c} \text{fail to determine} \\ \text{Spark}(F) \end{array}\right) &\leq \sum_{K=1}^M \binom{N}{K} \Pr\left(\mathcal{N}(H_{\mathcal{K}}^*) \cap \text{Span}(F_{\mathcal{K}'}) \neq \{0\}\right) \\ &\leq \sum_{K=1}^M \binom{N}{K} \frac{MK}{P} \\ &\leq \frac{M^3 2^N}{P}. \end{aligned}$$

Thus, to make the probability of failure $\leq \frac{1}{2}$, it suffices to have $P = M^3 2^{N+1}$.

In summary, we succeed in representing the original matroid with probability $\geq \frac{1}{2}$, and then we succeed in determining the spark of its representation with probability $\geq$

$\frac{1}{2}$. The probability of overall success is therefore $\geq \frac{1}{4}$. Since our success probability is, at worst, polynomially small, we can apply amplification to achieve an exponentially small error probability. $\square$

Our use of random linear projections in the above reduction to FULL SPARK is similar in spirit to Valiant and Vazirani's use of random hash functions in their reduction to UNIQUE SAT [VV86]. Since their randomized reduction is the canonical example thereof, we find our reduction to be particularly natural.

Chapter 5

Phase retrieval with polarization

5.1 Introduction

Given a collection of vectors $\Phi := \{\varphi_\ell\}_{\ell=1}^N \subseteq \mathbb{C}^M$ and a signal $x \in \mathbb{C}^M$, consider measurements of the form

$$z_\ell := |\langle x, \varphi_\ell \rangle|^2 + \nu_\ell, \quad (5.1.1)$$

where ν_ℓ is noise; we call these noisy *intensity measurements*. Several areas of imaging science, such as X-ray crystallography [Har93, MISE08, Mil90], diffraction imaging [BDP⁺07], astronomy [DF87] and optics [Wal63], use measurements of this form with the intent of reconstructing the original signal; this inverse problem is called *phase retrieval*. Note that in the measurement process (5.1.1), we inherently lose some information about x . Indeed, for every $\omega \in \mathbb{C}$ with $|\omega| = 1$, we see that x and ωx produce the same intensity measurements. Thus, the best one can hope to do with the intensity measurements of $x \in \mathbb{C}^M$ is reconstruct the class $[x] \in \mathbb{C}^M / \sim$, where $\sim$ is the equivalence relation of being identical up to a global phase factor.

In practice, phase retrieval falls short of determining the original signal up to global phase. First of all, the intensity measurement process that is used, say $\mathcal{A} : \mathbb{C}^M / \sim \rightarrow \mathbb{R}_{\geq 0}^N$ with $\mathcal{A}(x) = |\Phi^* x|^2$ (entrywise) and viewing $\Phi = [\varphi_1 \cdots \varphi_N]$, often

lacks injectivity, making it impossible to reconstruct uniquely. Moreover, the phase retrieval algorithms that are used in practice take alternating projections onto the column space of Φ^* (to bring phase to the measurements) and onto the nonconvex set of vectors y whose entry magnitudes match the intensity measurements $|\Phi^*x|^2$ (to maintain fidelity in the magnitudes) [Fie82, GS72, GJLM84]. The convergence of these algorithms is particularly sensitive to the choice of initial phases, and so even if the measurement design were injective, the phase retrieval procedure needs improvement.

These deficiencies have prompted two important lines of research in phase retrieval:

- (i) For which measurement designs Φ is $[x] \mapsto |\Phi^*x|^2$ injective?
- (ii) Given injectivity, how can $[x]$ be reconstructed stably and efficiently?

A first step toward solving (i) is determining how large N must be in order for $\mathcal{A}$ to be injective. It remains an open problem to find the smallest such N , but embedding results in differential geometry give that $N \geq (4 + o(1))M$ is necessary [AH59, HMW13]. As for sufficiency, Balan, Casazza and Edidin [BCE06] show that for almost every choice of Φ , $\mathcal{A}$ is injective whenever $N \geq 4M - 2$. However, Balan et al. do not suggest a signal reconstruction process in such cases. Even if a general process existed for the injective case, there is no guarantee it would be stable. Moreover, some instances of the phase retrieval problem are known to be NP-complete [SC91], and so any general reconstruction process is necessarily inefficient, assuming $P \neq NP$.

This leads one to attempt stable and efficient reconstruction with particular ensembles Φ . Until recently, this was only known to be possible in cases where $N = \Omega(M^2)$ [BBCE09]. By contrast, the state of the art comes from Candès, Strohmer and Voroninski [CSV11], who use semidefinite programming to stably re-

construct from $N = \mathcal{O}(M \log M)$ Gaussian-random measurements. There is other work along this vein [CESV13, DH12, WdM12] which also uses semidefinite programming and provides related guarantees. In theory, semidefinite programs are solved via interior point methods, whose iterations are often rather expensive. In practice, one is inclined to use faster numerical methods, but these lack performance guarantees.

In this paper, we propose an exchange of sorts: If you already have $\mathcal{O}(M \log M)$ Gaussian-random measurements, then we offer a faster reconstruction method with a stable performance guarantee, but at the price of $\mathcal{O}(M \log M)$ additional (non-adaptive) measurements. These new measurements are interferometry-inspired combinations of the originals, and the speedups gained in reconstruction come from our use of different spectral methods. To help motivate our measurement design and phase retrieval procedure, we start in the next section by considering the simpler, noiseless case. In this case, the success of our method follows from a neat trick involving the polarization identity along with some well-known results in the theory of expander graphs. In Section 3, we modify the method to obtain provable stability in the noisy case; here, we exploit some recent developments in spectral graph theory. We give concluding remarks in Section 4, and provide the more technical proofs in the appendix.

5.2 The noiseless case

In this section, we provide a new technique for phase retrieval. Here, we specifically address the noiseless case, in which ν_ℓ in (5.1.1) is zero for every $\ell = 1, \dots, N$; this case will give some intuition for a more stable version of our techniques, which we introduce in the next section. In the noiseless case, we will use on the order of the fewest measurements possible, namely $N = \mathcal{O}(M)$, where M is the dimension of the signal.

Before stating our measurement design and phase retrieval procedure, we motivate both with some discussion. Take a finite set V , and suppose we take intensity measurements of $x \in \mathbb{C}^M$ with a spanning set $\Phi_V := \{\varphi_i\}_{i \in V} \subseteq \mathbb{C}^M$. Again, we wish to recover x up to a global phase factor. Having $|\langle x, \varphi_i \rangle|$ for every $i \in V$, we claim it suffices to determine the relative phase between $\langle x, \varphi_i \rangle$ and $\langle x, \varphi_j \rangle$ for all pairs $i \neq j$. Indeed, if we had this information, we could arbitrarily assign some nonzero coefficient $c_i = |\langle x, \varphi_i \rangle|$ to have positive phase. If $\langle x, \varphi_j \rangle$ is also nonzero, then it has well-defined relative phase

$$\omega_{ij} := \left(\frac{\langle x, \varphi_i \rangle}{|\langle x, \varphi_i \rangle|} \right)^{-1} \frac{\langle x, \varphi_j \rangle}{|\langle x, \varphi_j \rangle|}, \quad (5.2.1)$$

which determines the coefficient by multiplication: $c_j = \omega_{ij} |\langle x, \varphi_j \rangle|$. Otherwise when $\langle x, \varphi_j \rangle = 0$, we naturally take $c_j = 0$, and for notational convenience, we arbitrarily take $\omega_{ij} = 1$. From here, the original signal's equivalence class $[x] \in \mathbb{C}^M / \sim$ can be identified by applying the canonical dual frame $\{\tilde{\varphi}_j\}_{j \in V}$, namely the Moore-Penrose pseudoinverse, of Φ_V :

$$\sum_{j \in V} c_j \tilde{\varphi}_j = \sum_{j \in V} \omega_{ij} |\langle x, \varphi_j \rangle| \tilde{\varphi}_j = \left(\frac{\langle x, \varphi_i \rangle}{|\langle x, \varphi_i \rangle|} \right)^{-1} \sum_{j \in V} \langle x, \varphi_j \rangle \tilde{\varphi}_j = \left(\frac{\langle x, \varphi_i \rangle}{|\langle x, \varphi_i \rangle|} \right)^{-1} x \in [x]. \quad (5.2.2)$$

Having established the utility of the relative phase between coefficients, we now seek some method of extracting this information. To this end, we turn to a special version of the polarization identity:

Lemma 5.2.1 (Mercedes-Benz Polarization Identity). *Take $\omega := e^{2\pi i/3}$. Then for any $a, b \in \mathbb{C}$,*

$$\bar{a}b = \frac{1}{3} \sum_{k=0}^2 \omega^k |a + \omega^{-k}b|^2. \quad (5.2.3)$$

Proof. We start by expanding the right-hand side of (5.2.3):

$$\text{RHS} := \frac{1}{3} \sum_{k=0}^2 \omega^k |a + \omega^{-k}b|^2 = \frac{1}{3} \sum_{k=0}^2 \omega^k (|a|^2 + 2\text{Re}(\omega^{-k}\bar{a}b) + |b|^2) = \frac{2}{3} \sum_{k=0}^2 \omega^k \text{Re}(\omega^{-k}\bar{a}b).$$

Multiplying, we find

$$\text{Re}(\omega^{-k}\bar{a}b) = \text{Re}(\omega^{-k})\text{Re}(\bar{a}b) - \text{Im}(\omega^{-k})\text{Im}(\bar{a}b) = \text{Re}(\omega^k)\text{Re}(\bar{a}b) + \text{Im}(\omega^k)\text{Im}(\bar{a}b).$$

We substitute this into our expression for RHS:

$$\begin{aligned} \text{Re}(\text{RHS}) &= \frac{2}{3} \left[\text{Re}(\bar{a}b) \sum_{k=0}^2 (\text{Re}(\omega^k))^2 + \text{Im}(\bar{a}b) \sum_{k=0}^2 \text{Re}(\omega^k)\text{Im}(\omega^k) \right], \\ \text{Im}(\text{RHS}) &= \frac{2}{3} \left[\text{Re}(\bar{a}b) \sum_{k=0}^2 \text{Re}(\omega^k)\text{Im}(\omega^k) + \text{Im}(\bar{a}b) \sum_{k=0}^2 (\text{Im}(\omega^k))^2 \right]. \end{aligned}$$

Finally, we apply the following easy-to-verify identities:

$$\sum_{k=0}^2 (\text{Re}(\omega^k))^2 = \sum_{k=0}^2 (\text{Im}(\omega^k))^2 = \frac{3}{2}, \quad \sum_{k=0}^2 \text{Re}(\omega^k)\text{Im}(\omega^k) = 0,$$

which yield $\text{RHS} = \bar{a}b$. □

The above polarization identity can also be proved by viewing $\{\omega^k\}_{k=0}^2$ as a Mercedes-Benz frame in $\mathbb{R}^2$ and $\frac{2}{3} \sum_{k=0}^2 \omega^k \text{Re}(\omega^{-k}u)$ as the corresponding reconstruction formula for $u \in \mathbb{C} = \mathbb{R}^2$. We can now use this polarization identity to determine relative phase (5.2.1):

$$\overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle = \frac{1}{3} \sum_{k=0}^2 \omega^k |\langle x, \varphi_i \rangle + \omega^{-k} \langle x, \varphi_j \rangle|^2 = \frac{1}{3} \sum_{k=0}^2 \omega^k |\langle x, \varphi_i + \omega^k \varphi_j \rangle|^2. \quad (5.2.4)$$

Thus, if in addition to Φ_V we measure with $\{\varphi_i + \omega^k \varphi_j\}_{k=0}^2$, we can use (5.2.4) to

determine $\overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle$ and then normalize to get the relative phase:

$$\omega_{ij} := \left(\frac{\langle x, \varphi_i \rangle}{|\langle x, \varphi_i \rangle|} \right)^{-1} \frac{\langle x, \varphi_j \rangle}{|\langle x, \varphi_j \rangle|} = \frac{\overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle}{|\langle x, \varphi_i \rangle \langle x, \varphi_j \rangle|}, \quad (5.2.5)$$

provided both $\langle x, \varphi_i \rangle$ and $\langle x, \varphi_j \rangle$ are nonzero. This idea of extracting information from interfering intensity measurements is not new; for example, [RSA⁺11] uses intensity measurements of multiple unknown signals and their interferences to recover them individually. To summarize our discussion of reconstructing a single signal, if we measure with Φ_V and $\{\varphi_i + \omega^k \varphi_j\}_{k=0}^2$ for every pair $i, j \in V$, then we can recover $[x]$. However, such a method uses $|V| + 3\binom{|V|}{2}$ measurements, and since Φ_V must span $\mathbb{C}^M$, we necessarily have $|V| \geq M$ and thus a total of $\Omega(M^2)$ measurements.

In pursuit of $\mathcal{O}(M)$ measurements, take some simple graph $G = (V, E)$, arbitrarily assign a direction to each edge, and only take measurements with Φ_V and $\Phi_E := \bigcup_{(i,j) \in E} \{\varphi_i + \omega^k \varphi_j\}_{k=0}^2$. To recover $[x]$, we again arbitrarily assign some nonzero vertex measurement to have positive phase, and then we propagate relative phase information along the edges by multiplication to determine the phase of the other vertex measurements relative to the original vertex measurement:

$$\omega_{ik} = \omega_{ij} \omega_{jk}. \quad (5.2.6)$$

However, if x is orthogonal to a given vertex vector, then that measurement is zero, and so relative phase information cannot propagate through the corresponding vertex; indeed, such orthogonality has the effect of removing the vertex from the graph, and for some graphs, this will prevent recovery. For example, if G is a star, then x could be orthogonal to the vector corresponding to the internal vertex, whose removal would render the remaining graph edgeless. That said, we should select Φ_V and G so as to minimize the impact of orthogonality with vertex vectors.

First, we can take Φ_V to be *full spark*, that is, Φ_V has the property that every

subcollection of M vectors spans. Full spark frames appear in a wide variety of applications. Explicit deterministic constructions of them are given in [ACM12, PK05]. For example, we can select the first M rows of the $|V| \times |V|$ discrete Fourier transform matrix, and take Φ_V to be the columns of the resulting $M \times |V|$ matrix; in this case, the fact that Φ_V is full spark follows from the Vandermonde determinant formula. In our application, Φ_V being full spark will be useful for two reasons. First, this implies that $x \neq 0$ is orthogonal to at most $M - 1$ members of Φ_V , thereby limiting the extent of x 's damage to our graph. Additionally, Φ_V being full spark frees us from requiring the graph to be connected after the removal of vertices; indeed, any remaining component of size M or more will correspond to a subcollection of Φ_V that spans, meaning it has a dual frame to reconstruct with. It remains to find a graph of $\mathcal{O}(M)$ vertices and edges that maintains a size- M component after the removal of any $M - 1$ vertices.

To this end, we consider a well-studied family of sparse graphs known as *expander graphs*. We choose these graphs for their notably strong connectivity properties. There is a combinatorial definition of expander graphs, but we will focus on the spectral definition. Given a d -regular graph G of n vertices, consider its adjacency matrix A , and define the *Laplacian* to be $L := I - \frac{1}{d}A$; if G were not regular, we would consider the diagonal matrix D of vertex degrees and define the Laplacian to be $L := I - D^{-1/2}AD^{-1/2}$. This is often called the *normalized* Laplacian in the literature, but we make no distinction here. We are particularly interested in the eigenvalues of the Laplacian: $0 = \lambda_1 \leq \dots \leq \lambda_n$. The second eigenvalue λ_2 of the Laplacian is called the *spectral gap* of the graph, and as we shall see, this value is particularly useful in evaluating the graph's connectivity. We say G has *expansion* λ if $\{\lambda_2, \dots, \lambda_n\} \subseteq [1 - \lambda, 1 + \lambda]$; note that since $1 - \lambda \leq \lambda_2$, small expansion implies large spectral gap. Furthermore, a family of d -regular graphs $\{G_i\}_{i=1}^\infty$ is a *spectral expander family* if there exists $c < 1$ such that every G_i has expansion $\lambda(G_i) \leq c$. Since d

is constant over an expander family, expanders with many vertices have particularly few edges. There are many results which describe the connectivity of expanders, but the following is particularly relevant to our application:

Lemma 5.2.2 (Spectral gap grants connectivity [HB05]). *Consider a d -regular graph G of n vertices with spectral gap λ_2 . For all $\varepsilon \leq \frac{\lambda_2}{6}$, removing any εdn edges from G results in a connected component of size $\geq (1 - \frac{2\varepsilon}{\lambda_2})n$.*

Note that removing εn vertices from a d -regular graph necessarily removes $\leq \varepsilon dn$ edges, and so this lemma directly applies. For our application, we want to guarantee that the removal of any $M - 1$ vertices maintains a size- M component. To do this, we will ensure both (i) $M - 1 \leq \varepsilon n$ and (ii) $M - 1 < (1 - \frac{2\varepsilon}{\lambda_2})n$, and then invoke the above lemma. Note that since $n \geq M \geq 2$,

$$\varepsilon \leq \frac{\lambda_2}{6} \leq \frac{\text{Tr}[L]}{6(n-1)} = \frac{n}{6(n-1)} \leq \frac{1}{3} < \frac{2}{3} \leq 1 - \frac{2\varepsilon}{\lambda_2},$$

where the last inequality is a rearrangement of $\varepsilon \leq \frac{\lambda_2}{6}$. Thus $\varepsilon n < (1 - \frac{2\varepsilon}{\lambda_2})n$, meaning (i) implies (ii), and so it suffices to have $M \leq \varepsilon n + 1$. Overall, we use the following criteria to pick our expander graph: Given the signal dimension M , use a d -regular graph $G = (V, E)$ of n vertices with spectral gap λ_2 such that $M \leq (\frac{\lambda_2}{6})n + 1$. Then by the previous discussion, the total number of measurements is $N = |V| + 3|E| = (\frac{3}{2}d + 1)n$.

Recall that we seek $N = \mathcal{O}(M)$ measurements. To minimize the redundancy $\frac{N}{M}$ for a fixed degree d , we would like a maximal spectral gap λ_2 , and it suffices to seek minimal spectral expansion λ . Spectral graph families known as *Ramanujan graphs* are asymptotically optimal in this sense; taking $\mathcal{G}_n^d$ to be the set of connected d -regular graphs with $\geq n$ vertices, Alon and Boppana (see [Alo86]) showed that for any fixed d ,

$$\lim_{n \rightarrow \infty} \inf_{G \in \mathcal{G}_n^d} \lambda(G) \geq \frac{2\sqrt{d-1}}{d},$$

while Ramanujan graphs are defined to have spectral expansion $\leq \frac{2\sqrt{d-1}}{d}$. To date, Ramanujan graphs have only been constructed for certain values of d . One important construction was given by Lubotzky, Phillips, and Sarnak [LPS88], which produces a Ramanujan family whenever $d - 1 \equiv 1 \pmod{4}$ is prime. Among these graphs, we get the smallest redundancy $\frac{N}{M}$ when $M = \lfloor (1 - \frac{2\sqrt{d-1}}{d})\frac{n}{6} + 1 \rfloor$ and $d = 6$:

$$\frac{N}{M} \leq \frac{(\frac{3}{2}d + 1)n}{(1 - \frac{2\sqrt{d-1}}{d})\frac{n}{6}} = 45(3 + \sqrt{5}) \approx 235.62.$$

Thus, in such cases, our techniques allow for phase retrieval with only $N \leq 236M$ measurements. However, the number of vertices in each Ramanujan graph from [LPS88] is of the form $q(q^2 - 1)$ or $\frac{q(q^2-1)}{2}$, where $q \equiv 1 \pmod{4}$ is prime, and so any bound on redundancy $\frac{N}{M}$ using these graphs will only be valid for particular values of M .

In order to get $N = \mathcal{O}(M)$ in general, we use the fact that random graphs are nearly Ramanujan with high probability. In particular, for every $\varepsilon > 0$ and even d , a random d -regular graph has spectral expansion $\lambda \leq \frac{2\sqrt{d-1}+\varepsilon}{d}$ with high probability as $n \rightarrow \infty$ [Fri08]. Thus, picking ε and d to satisfy $\frac{2\sqrt{d-1}+\varepsilon}{d} < 1$, we may take $M = \lfloor (1 - \frac{2\sqrt{d-1}+\varepsilon}{d})\frac{n}{6} + 1 \rfloor$ to get

$$\frac{N}{M} \leq \frac{(\frac{3}{2}d + 1)n}{(1 - \frac{2\sqrt{d-1}+\varepsilon}{d})\frac{n}{6}},$$

and this choice will satisfy $M \leq (\frac{\lambda_2}{6})n + 1$ with high probability. To see how small this redundancy is, note that taking $\varepsilon = 0.1$ and $d = 8$ gives $N \leq 240M$. While the desired expansion properties of a random graph are only present with high probability, estimating the spectral gap is inexpensive, and so it is computationally feasible to verify whether a randomly drawn graph is good enough. Moreover, n can be any sufficiently large integer, and so the above bound is valid for all sufficiently large M , i.e., our procedure can perform phase retrieval with $N = \mathcal{O}(M)$ measurements in general.

Combining this with the above discussion, we have the following measurement design and phase retrieval procedure:

Measurement design (noiseless case)

- Fix d even and $\varepsilon > 0$.
- Given M , pick some d -regular graph $G = (V, E)$ with spectral gap $\lambda_2 \geq \lambda' := 1 - \frac{2\sqrt{d-1}+\varepsilon}{d}$ and $|V| = \lceil \frac{6}{\lambda'}(M-1) \rceil$, and arbitrarily direct the edges.
- Design the measurements $\Phi := \Phi_V \cup \Phi_E$ by taking $\Phi_V := \{\varphi_i\}_{i \in V} \subseteq \mathbb{C}^M$ to be full spark and $\Phi_E := \bigcup_{(i,j) \in E} \{\varphi_i + \omega^k \varphi_j\}_{k=0}^3$.

Phase retrieval procedure (noiseless case)

- Given $\{|\langle x, \varphi \rangle|^2\}_{\varphi \in \Phi}$, delete the vertices $i \in V$ with $|\langle x, \varphi_i \rangle|^2 = 0$.
- In the remaining induced subgraph, find a connected component of $\geq M$ vertices V' .
- Pick a vertex in V' to have positive phase and propagate/multiply relative phases (5.2.6), which are calculated by normalizing (5.2.4), see (5.2.5).
- Having $\{\langle x, \varphi_i \rangle\}_{i \in V'}$ up to a global phase factor, find the least-squares estimate of $[x]$ by applying the Moore-Penrose pseudoinverse of $\{\varphi_i\}_{i \in V'}$, see (5.2.2).

Note that this phase retrieval procedure is particularly fast. Indeed, if we use $E \subseteq V^2$ to store G , then we can delete vertices $i \in V$ with $|\langle x, \varphi_i \rangle|^2 = 0$ by deleting the edges for which (5.2.4) is zero, which takes $\mathcal{O}(|E|)$ time. Next, if the members of E are ordered lexicographically, the remaining subgraph can be easily partitioned into connected components in $\mathcal{O}(|E|)$ time by collecting edges with common vertices, and then propagating relative phase in the largest component is performed in $\mathcal{O}(|E|)$ time using a depth- or breadth-first search. Overall, we only use $\mathcal{O}(M)$ time before the final least-squares step of the phase retrieval procedure, which happens to be the

bottleneck, depending on the subcollection $\Phi_{V'}$. In general, we can find the least-squares estimate in $\mathcal{O}(M^3)$ time using Gaussian elimination, but if $\Phi_{V'}$ has special structure (e.g., it is a submatrix of the discrete Fourier transform matrix), then one might exploit that structure to gain speedups (e.g., use the fast Fourier transform in conjunction with an iterative method). Regardless, our procedure reduces the nonlinear phase retrieval problem to the much simpler problem of solving an overdetermined *linear* system.

While this measurement design and phase retrieval procedure is particularly efficient, it certainly lacks stability. Perhaps most notably, we have not imposed anything on Φ_V that guarantees stability with inverting $\Phi_{V'}$; indeed, we have merely enforced linear independence between vectors, while stability will require well-conditioning. Another noteworthy source of instability is our method of phase propagation, which naturally accumulates error; it would be better if the relative phases were combined using a more democratic process that encourages noise cancellation. In the next section, we will address these concerns (and others) and modify our procedure accordingly; the revised procedure will be stable, but at the price of a log factor in the number of measurements: $N = \mathcal{O}(M \log M)$. As we mention in the concluding remarks, we do not think this log factor is necessary, but we leave this pursuit for future work.

5.3 The noisy case

In this section, we consider a noise-robust version of the measurement design and phase retrieval procedure of the previous section. In the end, the measurement design will be nearly identical: vertex measurements will be independent complex Gaussian vectors (thereby being full spark with probability 1), and the edge measurements will be the same sort of linear combinations of vertex measurements. Our use of

randomness in this version will enable the vertex measurements to simultaneously satisfy two important conditions with high probability: *projective uniformity with noise* and *numerical erasure robustness*. Before defining these conditions, we motivate them by considering a noisy version of our phase retrieval procedure.

Recall that our noiseless procedure starts by removing the vertices $i \in V$ for which $|\langle x, \varphi_i \rangle|^2 = 0$. Indeed, since we plan to propagate relative phase information along edges, these 0-vertices are of no use, as relative phase with these vertices is not well defined. Since we calculate relative phase by normalizing (5.2.4), we see that relative phase is sensitive to perturbations when (5.2.4) is small, meaning either $\langle x, \varphi_i \rangle$ or $\langle x, \varphi_j \rangle$ is small. As such, while 0-vertices provide no relative phase information in the noiseless case, small vertices provide *unreliable* information in the noisy case, and so we wish to remove them accordingly. However, we also want to ensure that there are only a few small vertices. In the noiseless case, we limit the number of 0-vertices by using a full spark frame; in the noisy case, we make use of a new concept we call *projective uniformity*:

Definition 5.3.1. The α -projective uniformity of $\Phi = \{\varphi_i\}_{i=1}^n \subseteq \mathbb{C}^M$ is given by

$$\text{PU}(\Phi; \alpha) = \min_{\substack{x \in \mathbb{C}^M \\ |x|=1}} \max_{\substack{\mathcal{I} \subseteq \{1, \dots, n\} \\ |\mathcal{I}| \geq \alpha n}} \min_{i \in \mathcal{I}} |\langle x, \varphi_i \rangle|^2.$$

In words, projective uniformity gives the following guarantee: For every unit-norm signal x , there exists a collection of vertices $\mathcal{I} \subseteq V$ of size at least $\alpha |V|$ such that $|\langle x, \varphi_i \rangle|^2 \geq \text{PU}(\Phi_V; \alpha)$ for every $i \in \mathcal{I}$. As such, projective uniformity effectively limits the total number of small vertices possible, at least before the measurements are corrupted by noise. However, the phase retrieval algorithm will only have access to noisy versions of the measurements, and so we must account for this subtlety in our procedure. In an effort to isolate the reliable pieces of relative phase information, we will remove the vertices corresponding to small noisy edge combinations (5.2.4):

Algorithm 5.3.2 (Pruning for reliability). *Given a graph $G = (V, E)$ and a function $f: E \rightarrow \mathbb{R}$ such that $f(i, j) = \left| \overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle + \varepsilon_{ij} \right|$, initialize $H \leftarrow G$. Do the following procedure $\lfloor (1 - \alpha) |V| \rfloor$ times: Find the minimizer $(i, j) \in E$ of f , and update $H \leftarrow H \setminus \{i, j\}$. When done, output H .*

We now explain why only reliable pieces of relative phase information will remain after running the above algorithm, provided Φ_V has sufficient projective uniformity. The main idea is captured in the following:

Lemma 5.3.3. *Define $|\theta|_{\mathbb{T}} := \min_{k \in \mathbb{Z}} |\theta - 2\pi k|$ for all angles $\theta \in \mathbb{R}/2\pi\mathbb{Z}$. Then for any $z, \varepsilon \in \mathbb{C}$,*

$$|\arg(z + \varepsilon) - \arg(z)|_{\mathbb{T}} \leq \pi \frac{|\varepsilon|}{|z|}.$$

Proof. If $|\varepsilon| \geq |z|$, the result is trivial. Suppose $|\varepsilon| < |z|$, and consider the triangle whose vertices are 0 , z , and $z + \varepsilon$. Denoting the angle at 0 by $\theta = |\arg(z + \varepsilon) - \arg(z)|_{\mathbb{T}}$ and the angle at $z + \varepsilon$ by α , the law of sines gives

$$\frac{\sin \theta}{|\varepsilon|} = \frac{\sin \alpha}{|z|} \leq \frac{1}{|z|}. \quad (5.3.1)$$

Next, the law of cosines gives

$$2|z||z + \varepsilon| \cos \theta = |z|^2 + |z + \varepsilon|^2 - |\varepsilon|^2 > |z + \varepsilon|^2 > 0,$$

and so $\cos \theta > 0$, i.e., $\theta \in [0, \frac{\pi}{2})$. Finally, by the concavity of $\sin(\cdot)$ and then (5.3.1), we conclude that

$$\frac{2}{\pi} \theta \leq \sin \theta \leq \frac{|\varepsilon|}{|z|},$$

which implies the result. $\square$

By taking $z = \overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle + \varepsilon_{ij}$ and $\varepsilon = -\varepsilon_{ij}$, we can use this lemma to bound the relative phase error we incur when normalizing z . In fact, consider the minimum

of f when Algorithm 5.3.2 is complete. Since the algorithm deletes vertices from G according to the input signal x , this minimum will vary with x ; let PUN denote the smallest possible minimum value. Then the relative phase error incurred with $(i, j) \in E$ is no more than $\pi |\varepsilon_{ij}| / \text{PUN}$, regardless of the signal measured. Indeed, our use of *projective uniformity with noise* (i.e., PUN) is intended to bound the instability that comes with normalizing small values of (5.2.4). As we will show in the appendix, PUN can be bounded below by using the projective uniformity of Φ_V , and furthermore, a complex Gaussian Φ_V has projective uniformity with overwhelming probability.

After applying Algorithm 5.3.2, our graph will have slightly fewer vertices, but the remaining edges will correspond to reliable pieces of relative phase information. Recall that we plan to use this information on the edges to determine phases for the vertices, and we want to do this in a stable way. To understand when this is even possible, we first consider a few simple scenarios. Suppose that after removing vertices with Algorithm 5.3.2, the graph has a vertex of degree 0. Then we have no information about the phase of this vertex, and it should be removed accordingly. For a less extreme scenario, suppose the vertex has degree 1. Then any noise in the corresponding edge measurement would be passed directly to the vertex, which inherently lacks stability compared to the noise cancellation that would come with more edges. More generally, if the graph has a cut vertex (e.g., the neighbor of a degree-1 vertex), then we would need to rely on the correctness of this lone vertex to ensure consistency between the parts of the graph it connects—this scenario is also rather unstable. After considering these examples, it makes intuitive sense that stability necessitates a high level of connectivity in the graph, regardless of the algorithm used to extrapolate the vertex phases.

As such, we seek to remove a small proportion of vertices so that the remaining graph is very connected, i.e., has large spectral gap. To do this, we will iteratively

remove sets of vertices that are poorly connected to the rest of the graph. These sets will be identified using *spectral clustering*, a process which is strongly motivated by an inequality in Riemannian geometry by Cheeger [Che70] and which has performance guarantees originating with Alon [Alo86, AM85]. The main idea of spectral clustering follows the intuition that a random walk on a graph tends to be trapped in sections of the graph which have few connections to the rest of the vertices. Moreover, the second eigenvector of the corresponding stochastic matrix tends to identify these sections.

Algorithm 5.3.4 (Spectral clustering). *Given a graph, compute the eigenvector u corresponding to the second eigenvalue of its Laplacian. Let S_i denote the vertices corresponding to the i smallest entries of $D^{-1/2}u$. Minimize*

$$h(S_i) = \frac{E(S_i, S_i^c)}{\min\{\text{vol}(S_i), \text{vol}(S_i^c)\}},$$

where $E(S, T)$ denotes the number of edges between S and T and

$$\text{vol}(S) = \sum_{v \in S} \deg(v).$$

Output the minimizer S_k or its complement, whichever is smaller in size.

Note that when implementing spectral clustering, the values of $E(S_i, S_i^c)$, $\text{vol}(S_i)$ and $\text{vol}(S_i^c)$ can be cheaply updated from their $(i - 1)$ st values since S_i differs from S_{i-1} in only one vertex. As such, the bottleneck in spectral clustering is computing an eigenvector. For our application, we will iteratively apply spectral clustering to identify small collections of vertices which are poorly connected to the rest of the graph and then remove them to enhance connectivity:

Algorithm 5.3.5 (Pruning for connectivity). *Given a graph G and a threshold $\tau > 0$, initialize $H \leftarrow G$. While the spectral gap of H is smaller than τ , perform spectral clustering (Algorithm 5.3.4) to identify a small set of vertices S , and update $H \leftarrow$*

$H \setminus S$. Once the spectral gap of H is at least τ , output H .

In the appendix, we show that for a particular choice of threshold τ , Algorithm 5.3.5 recovers a level of connectivity that may have been lost when pruning for reliability in Algorithm 5.3.2, and it does so by removing only a small proportion of the vertices.

At this point, we have pruned our graph so that the measured relative phases are reliable and the vertex phases can be stably reconstructed. Now we seek an efficient method to reconstruct these vertex phases from the measured relative phases. Before devising such a method, we first organize the information we have into a matrix. Given the graph output $G' = (V', E')$ of Algorithm 5.3.5, we take A_1 to be the $|V'| \times |V'|$ weighted adjacency matrix with normalized noisy versions of (5.2.4):

$$A_1[i, j] = \frac{\overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle + \varepsilon_{ij}}{\left| \overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle + \varepsilon_{ij} \right|} \quad (5.3.2)$$

whenever $\{i, j\} \in E'$, and otherwise $A_1[i, j] = 0$. Unlike the noiseless case, here, we account for both directions (i, j) and (j, i) whenever $\{i, j\} \in E'$, with the understanding that $\varepsilon_{ji} = \overline{\varepsilon_{ij}}$; this will simplify our analysis since this makes A_1 self-adjoint. Considering $A_1[i, j]$ is an approximation of the relative phase $\omega_i^{-1} \omega_j$, it seems reasonable to extrapolate the vertex phases $\omega := \{\omega_i\}_{i \in V'}$ from A_1 by minimizing the following quantity:

$$\sum_{\{i, j\} \in E'} |\omega_j - A_1[i, j] \omega_i|^2 = \sum_{\{i, j\} \in E'} \left(|\omega_j|^2 - 2 \operatorname{Re} \overline{\omega_j} A_1[i, j] \omega_i + |A_1[i, j] \omega_i|^2 \right) = \omega^* (D - A_1) \omega,$$

where D is the diagonal matrix of vertex degrees. Dividing by $\operatorname{vol}(G') = \omega^* D \omega$, which does not vary with ω , this is equivalent to minimizing

$$\frac{\omega^* (D - A_1) \omega}{\omega^* D \omega} = \frac{(D^{1/2} \omega)^* (I - D^{-1/2} A_1 D^{-1/2}) (D^{1/2} \omega)}{|D^{1/2} \omega|^2} \geq \lambda_1 (I - D^{-1/2} A_1 D^{-1/2}).$$

To be clear, the right-hand side above is the first eigenvalue of $L_1 := I - D^{-1/2}A_1D^{-1/2}$, which we call the *connection Laplacian*; note that this bears some resemblance to the Laplacian defined in the previous section. In minimizing the above quantity, it makes sense to consider the eigenvector u corresponding to the smallest eigenvalue of L_1 , but we require each coordinate of $D^{-1/2}u$ to have unit modulus. Provided u has no entries which are zero, we can normalize the entries to form an estimate of ω , and as we show in the appendix (using results from [BSS12]), this estimate is stable provided the spectral gap of G' is sufficiently large. This spectral method is known in the literature as *angular synchronization* [Sin11], and we summarize the procedure with the following:

Algorithm 5.3.6 (Angular synchronization). *Given a graph $G' = (V', E')$ and noisy versions of (5.2.4) for every $\{i, j\} \in E'$, construct a weighted adjacency matrix A_1 according to (5.3.2) whenever $\{i, j\} \in E'$, and otherwise $A_1[i, j] = 0$. Take D to be the diagonal matrix of vertex degrees and define the connection Laplacian $L_1 := I - D^{-1/2}A_1D^{-1/2}$. Compute the eigenvector corresponding to the smallest eigenvalue of L_1 , and output normalized versions of its coordinates.*

To reiterate, Algorithm 5.3.6 will produce estimates for the phases of the inner products $\{\langle x, \varphi_i \rangle\}_{i \in V'}$. Also, we can take square roots of the vertex measurements $\{|\langle x, \varphi_i \rangle|^2 + \nu_i\}_{i \in V'}$ to estimate $\{|\langle x, \varphi_i \rangle|\}_{i \in V'}$. Then we can combine these to estimate $\{\langle x, \varphi_i \rangle\}_{i \in V'}$. However, note that the largest of these inner products will be most susceptible to noise in the corresponding phase estimate. As such, we remove a small fraction of these largest vertices so that the final collection of vertices V'' has size $\kappa|V|$, where V was the original vertex set, and κ is sufficiently close to 1.

Now that we have estimated the phases of $\{\langle x, \varphi_i \rangle\}_{i \in V''}$, we wish to reconstruct x by applying the Moore-Penrose pseudoinverse of $\{\varphi_i\}_{i \in V''}$. However, since V'' is likely a strict subset of V , it can be difficult in general to predict how stable the pseudoinverse will be. Fortunately, a recent theory of *numerically erasure-robust frames*

(*NERFs*) makes this prediction possible: If the members of Φ_V are independent Gaussian vectors, then with high probability, every submatrix of columns $\Phi_{V''}$ with $\kappa = |V''| / |V|$ sufficiently large has a stable pseudoinverse [FM12]. This concludes the phase retrieval procedure, briefly outlined below together with the measurement design.

Measurement design (noisy case)

- Fix d even and $\varepsilon > 0$.
- Given M , pick some d -regular graph $G = (V, E)$ with spectral gap $\lambda_2 \geq \lambda' := 1 - \frac{2\sqrt{d-1}+\varepsilon}{d}$ and $|V| = cM \log M$ for c sufficiently large, and arbitrarily direct the edges.
- Design the measurements $\Phi := \Phi_V \cup \Phi_E$ by taking $\Phi_V := \{\varphi_i\}_{i \in V} \subseteq \mathbb{C}^M$ to have independent entries with distribution $\mathbb{CN}(0, \frac{1}{M})$ and $\Phi_E := \bigcup_{(i,j) \in E} \{\varphi_i + \omega^k \varphi_j\}_{k=0}^3$.

Phase retrieval procedure (noisy case)

- Given $\{|\langle x, \varphi_\ell \rangle|^2 + \nu_\ell\}_{\ell=1}^N$, prune the graph G , keeping only reliable vertices (Algorithm 5.3.2).
- Prune the remaining induced subgraph for connectivity, producing the vertex set V' (Algorithm 5.3.5).
- Estimate the phases of the vertex measurements using angular synchronization (Algorithm 5.3.6).
- Remove the vertices with the largest measurements, keeping only $|V''| = \kappa |V|$.
- Having estimates for

$$\{\langle x, \varphi_i \rangle\}_{i \in V''}$$

up to a global phase factor, find the least-squares estimate of $[x]$ by applying the Moore-Penrose pseudoinverse of $\{\varphi_i\}_{i \in V''}$, see (5.2.2).

Having established our measurement design and phase retrieval procedure for the noisy case, we now present the following guarantee of stable performance:

Theorem 5.3.7. *Pick $N \sim CM \log M$ with C sufficiently large, and take $\{\varphi_\ell\}_{\ell=1}^N = \Phi_V \cup \Phi_E$ defined in the measurement design above. Then there exist constants $C', K > 0$ such that the following guarantee holds for all $x \in \mathbb{C}^M$ with overwhelming probability: Consider measurements of the form*

$$z_\ell := |\langle x, \varphi_\ell \rangle|^2 + \nu_\ell.$$

If the noise-to-signal ratio satisfies $\text{NSR} := \frac{|\nu|}{|x|^2} \leq \frac{C'}{\sqrt{M}}$, then the phase retrieval procedure above produces an estimate $\tilde{x}$ from $\{z_\ell\}_{\ell=1}^N$ with squared relative error

$$\frac{|\tilde{x} - e^{i\theta}x|^2}{|x|^2} \leq K \sqrt{\frac{M}{\log M}} \text{NSR}$$

for some phase $\theta \in [0, 2\pi)$.

The interested reader is directed to the appendix for a proof of this guarantee. Before concluding this section, we briefly evaluate the result. Note that the norms of the φ_ℓ 's tend to be $\mathcal{O}(1)$, and so the noiseless measurements $|\langle x, \varphi_\ell \rangle|^2$ tend to be of size $\mathcal{O}(|x|^2/M)$. Also, in the worst-case scenario, the noise annihilates our measurements $\nu_\ell = -|\langle x, \varphi_\ell \rangle|^2$, rendering the signal x unrecoverable; in this case, $|\nu| = \mathcal{O}(|x|^2 \sqrt{(\log M)/M})$ since $N = CM \log M$. In other words, if we allowed the noise-to-signal ratio to scale slightly larger than $C'/\sqrt{M}$ (i.e., by a log factor), then it would be impossible to perform phase retrieval in the worst case. As such, the above guarantee is optimal in some sense. Furthermore, since $\sqrt{M/\log M} \text{NSR} =$

$\mathcal{O}(1/\sqrt{\log M})$ by assumption, the result indicates that our phase retrieval process exhibits more stability as M grows large.

5.4 Concluding remarks

This paper provides a new way to perform phase retrieval, and our main result (Theorem 5.3.7) shows that our method is stable. In comparison to the state of the art, namely, the work of Candès, Strohmer and Voroninski [CSV11], our stability results mainly differ in how we choose to scale the measurement vectors. Indeed, the measurement vectors of the present paper tend to have norm $\mathcal{O}(1)$, whereas the measurement vectors of Candès et al. are all scaled to have norm $\sqrt{M}$; considering the statement of Theorem 5.3.7 is riddled with square roots of M , either choice of scaling is arguably natural.

One might feel that our phase retrieval algorithms are slightly unsatisfying because we perform hard thresholds to remove vertices according to how small or large the corresponding measurements are. Alternatively, there could very well be a way to more smoothly weight these measurements according to our confidence in them, and such weightings are already accounted for in the theory of angular synchronization [BSS12]. However, we decided to use hard thresholds because they greatly simplify the analysis of projective uniformity (though the analysis is still rather technical).

While the worst-case analysis we provide here is useful in many applications (and enables a comparison with the worst-case stability results of [CSV11]), stochastic noise is a more appropriate model in other applications. We believe that the phase retrieval procedure of this paper will perform substantially better in the average case, but we leave this analysis for future work. Also, a notable distinction between our measurement designs in the noiseless and noisy cases is the presence of a log factor in the number of measurements used. However, we believe this factor is an artifact

of our current analysis, and we intend to remove it in the future.

5.5 Appendix

5.5.1 Graph pruning

This section proves the following guarantee:

Theorem 5.5.1. *Take proportions $p \geq q \geq \frac{2}{3}$, and consider a regular graph $G = (V, E)$ with spectral gap $\lambda_2 > g(p, q) := 1 - 2(q(1 - q) - (1 - p))$. After Algorithm 5.3.2 removes at most $(1 - p)|V|$ vertices from G , then setting $\tau = \frac{1}{8}(\lambda_2 - g(p, q))^2$, Algorithm 5.3.5 outputs a subgraph with at least $q|V|$ vertices.*

To prove this theorem, we will apply a graph version of the Cheeger inequality, which provides a guarantee for Algorithm 5.3.4:

Theorem 5.5.2 (Constructive Cheeger inequality [Chu10]). *Consider a graph $G = (V, E)$ with spectral gap λ_2 . Then Algorithm 5.3.4 outputs a set of vertices S such that $h(S) \leq \sqrt{2\lambda_2}$.*

Proof of Theorem 5.5.1. First, Algorithm 5.3.2 removes a set of vertices, which we denote by S_0 . In applying Algorithm 5.3.5, the i th step of the while loop removes another set of vertices S_i . We claim this while loop will end with $|\bigcup_{i \geq 0} S_i| < (1 - q)|V|$. Supposing to the contrary, consider the first k for which $S := \bigcup_{i=0}^k S_i$ has at least $(1 - q)|V|$ vertices. Then since each iteration of the while loop removes at most half of the remaining vertices, we have $(1 - q)|V| \leq |S| \leq (1 - \frac{q}{2})|V|$.

To derive a contradiction, we will find incompatible upper and lower bounds on $E(S, S^c)$. For the upper bound, we apply the fact that G is d -regular along with the definition of $h(S_i)$ to get

$$E(S, S^c) = E(S_0, S^c) + \sum_{i=1}^k E(S_i, S^c) \leq \text{vol}(S_0) + \sum_{i=1}^k E(S_i, S_i^c) \leq d|S_0| + \sum_{i=1}^k h(S_i) \text{vol}(S_i).$$

Next, Theorem 5.5.2 bounds each $h(S_i)$ in terms of the spectral gap the remaining graph, which is necessarily less than τ by the condition of the while loop. Thus, we continue:

$$E(S, S^c) < d|S_0| + \sum_{i=1}^k \sqrt{2\tau}d|S_i| \leq d(1-p)|V| + \sqrt{2\tau}d|S| \leq d|V| \left((1-p) + \sqrt{2\tau}(1-\frac{q}{2}) \right). \quad (5.5.1)$$

For the lower bound, we use the expander mixing lemma, which says

$$\left| E(S, S^c) - \frac{d}{|V|} |S| |S^c| \right| \leq d(1 - \lambda_2) \sqrt{|S| |S^c|}.$$

Since the function $x \mapsto x(1-x)$ is concave down, we know the minimum subject to $1-q \leq x \leq 1-\frac{q}{2}$ is achieved at an endpoint of this interval. Thus, evaluating the function at $|S|/|V|$ gives $|S| |S^c| / |V|^2 \geq \min\{q(1-q), \frac{q}{2}(1-\frac{q}{2})\} = q(1-q)$, where the last step follows from the fact that $q \geq \frac{2}{3}$. Applying this and $|S| |S^c| \leq |V|^2/4$ to the expander mixing lemma then gives

$$E(S, S^c) \geq \frac{d}{|V|} |S| |S^c| - d(1-\lambda_2) \sqrt{|S| |S^c|} \geq d|V| q(1-q) - d(1-\lambda_2) \frac{|V|}{2} = d|V| \left(q(1-q) - \frac{1-\lambda_2}{2} \right). \quad (5.5.2)$$

Finally, we combine (5.5.1) and (5.5.2) to get

$$q(1-q) - \frac{1-\lambda_2}{2} < (1-p) + \sqrt{2\tau}(1-\frac{q}{2}),$$

which, as substitution reveals, contradicts our choice for τ . □

5.5.2 Angular synchronization

This section proves the following guarantee:

Theorem 5.5.3. *Consider a graph $G = (V, E)$ with spectral gap $\lambda_2 > 0$, and define $|\theta|_{\mathbb{T}} := \min_{k \in \mathbb{Z}} |\theta - 2\pi k|$ for all angles $\theta \in \mathbb{R}/2\pi\mathbb{Z}$. Given the weighted adjacency*

matrix A_1 (5.3.2), then Algorithm 5.3.6 outputs $u \in \mathbb{C}^{|V|}$ with unit-modulus entries such that, for some phase $\theta \in \mathbb{R}/2\pi\mathbb{Z}$,

$$\sum_{i \in V} |\arg(u_i) - \arg(\langle x, \varphi_i \rangle) - \theta|_{\mathbb{T}}^2 \leq \frac{C |\varepsilon|^2}{\lambda_2^2 P^2},$$

where $P := \min_{\{i,j\} \in E} \left| \overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle + \varepsilon_{ij} \right|$ and C is a universal constant.

In terms of phases, we are interested in reconstructing $\omega^* : V \rightarrow \mathbb{R}/2\pi\mathbb{Z}$ such that $\omega_i^* = \arg(\langle x, \varphi_i \rangle)$ for every $i \in V$. For some graph $G = (V, E)$, we are given a weighted adjacency matrix A_1 (5.3.2), which encodes edge measurements:

$$\rho_{ij} := \omega_j^* - \omega_i^* + e_{ij}$$

for every $(i, j) \in E$; here, $e_{ij} \in \mathbb{R}/2\pi\mathbb{Z}$ is angular noise, and we note that $e_{ji} = -e_{ij}$. We measure the size of $e = \{e_{ij}\}_{\{i,j\} \in E}$ in terms of its components: $|e|_{\mathbb{T}}^2 := \sum_{\{i,j\} \in E} |e_{ij}|_{\mathbb{T}}^2$.

We will prove Theorem 5.5.3 using a recent Cheeger inequality for the connection Laplacian [BSS12]. In effect, this result provides a guarantee for Algorithm 5.3.6 in terms of a certain objective function. To be precise, given $\omega : V \rightarrow \mathbb{R}/2\pi\mathbb{Z}$, we define

$$\eta(\omega) = \sum_{\{i,j\} \in E} \left| e^{i((\omega_j - \omega_i) - \rho_{ij})} - 1 \right|^2.$$

Note that one way to reconstruct ω^* is to minimize this quantity. Indeed, $\eta(\omega)$ is small when the angular differences $\omega_j - \omega_i$ are close to the measured differences ρ_{ij} , and in the noiseless case, $\eta(\omega) = 0$ precisely when $\omega = \omega^*$, provided the graph G is connected. In terms of this objective function, the following guarantee ensures that the output of Algorithm 5.3.6 is no worse than a constant multiple of optimal:

Theorem 5.5.4 (Cheeger inequality for the connection Laplacian [BSS12]). *Consider*

a graph $G = (V, E)$ with spectral gap $\lambda_2 > 0$. Given the weighted adjacency matrix A_1 (5.3.2), which encodes edge measurements $\rho : E \rightarrow \mathbb{R}/2\pi\mathbb{Z}$, then Algorithm 5.3.6 outputs $u \in \mathbb{C}^{|V|}$, which encodes $\hat{\omega} : V \rightarrow \mathbb{R}/2\pi\mathbb{Z}$ such that

$$\eta(\hat{\omega}) \leq \frac{C'}{\lambda_2} \min_{\omega: V \rightarrow \mathbb{T}} \eta(\omega), \quad (5.5.3)$$

where C' is a universal constant.

We start with two lemmas:

Lemma 5.5.5. *For every $a, b \in \mathbb{R}/2\pi\mathbb{Z}$, we have $\frac{1}{2} |a|_{\mathbb{T}}^2 - |b|_{\mathbb{T}}^2 \leq |a - b|_{\mathbb{T}}^2$.*

Proof. By abuse of notation, we identify a and b with their coset representatives in $[-\pi, \pi)$. Since

$$0 \leq \frac{1}{2}(a - 2b)^2 = (a - b)^2 - (\frac{1}{2}a^2 - b^2),$$

then rearranging gives the following inequality:

$$\frac{1}{2} |a|_{\mathbb{T}}^2 - |b|_{\mathbb{T}}^2 = \frac{1}{2}a^2 - b^2 \leq (a - b)^2 = |a - b|_{\mathbb{T}}^2,$$

where the last equality requires $a - b \in [-\pi, \pi]$. Otherwise, we note that

$$0 \leq \frac{1}{2}(a - 2b \pm 2\pi)^2 + 2\pi(\pi \pm a) = (a - b \pm 2\pi)^2 - (\frac{1}{2}a^2 - b^2),$$

and so rearranging gives

$$\frac{1}{2} |a|_{\mathbb{T}}^2 - |b|_{\mathbb{T}}^2 = \frac{1}{2}a^2 - b^2 \leq (a - b \pm 2\pi)^2,$$

one of which equals $|a - b|_{\mathbb{T}}^2$ depending on whether $a - b < -\pi$ or $a - b > \pi$. $\square$

Lemma 5.5.6. *For every $a, b \in \mathbb{C}$ with $|a| = 1$, we have $\left|a - \frac{b}{|b|}\right| \leq 2|a - b|$.*

Proof. The reverse triangle inequality gives $\left| \frac{b}{|b|} - b \right| = |1 - |b|| = ||a| - |b|| \leq |a - b|$, and so the triangle inequality gives $\left| a - \frac{b}{|b|} \right| \leq |a - b| + \left| b - \frac{b}{|b|} \right| \leq 2|a - b|$. $\square$

Proof of Theorem 5.5.3. Identifying $\theta \in \mathbb{R}/2\pi\mathbb{Z}$ with its coset representative in $[-\pi, \pi)$, we have $|\theta|_{\mathbb{T}} = |\theta|$, and a double-angle formula gives

$$|e^{i\theta} - 1| = 2 \sin \frac{|\theta|}{2}.$$

From these, it follows that

$$\frac{2}{\pi} |\theta|_{\mathbb{T}} \leq |e^{i\theta} - 1| \leq |\theta|_{\mathbb{T}}. \quad (5.5.4)$$

This relationship will allow us to apply Theorem 5.5.4. To this end, for notational convenience, we define $\gamma_i := \arg(u_i) - \arg(\langle x, \varphi_i \rangle)$ for every $i \in V$. The right-hand inequality of (5.5.4) and Lemma 5.5.5 together give

$$\frac{1}{2} \sum_{\{i,j\} \in E} |e^{i(\gamma_j - \gamma_i)} - 1|^2 - |e|_{\mathbb{T}}^2 \leq \sum_{\{i,j\} \in E} \left(\frac{1}{2} |\gamma_j - \gamma_i|_{\mathbb{T}}^2 - |e_{ij}|_{\mathbb{T}}^2 \right) \leq \sum_{\{i,j\} \in E} |\gamma_j - \gamma_i - e_{ij}|_{\mathbb{T}}^2. \quad (5.5.5)$$

Denoting $\hat{\omega}_i := \arg(u_i)$ and $\omega_i^* := \arg(\langle x, \varphi_i \rangle)$, then the definition of η gives

$$\eta(\hat{\omega}) = \sum_{\{i,j\} \in E} |e^{i(\gamma_j - \gamma_i - e_{ij})} - 1|^2.$$

With this, we continue (5.5.5) by applying the left-hand inequality of (5.5.4):

$$\frac{1}{2} \sum_{\{i,j\} \in E} |e^{i(\gamma_j - \gamma_i)} - 1|^2 - |e|_{\mathbb{T}}^2 \leq \frac{\pi^2}{4} \eta(\hat{\omega}) \leq \frac{\pi^2}{4} \frac{C'}{\lambda_2} \min_{\omega: V \rightarrow \mathbb{T}} \eta(\omega) \leq \frac{\pi^2}{4} \frac{C'}{\lambda_2} \eta(\omega^*), \quad (5.5.6)$$

where the second inequality follows from Theorem 5.5.4. Furthermore, the right-hand

inequality of (5.5.4) gives

$$\eta(\omega^*) = \sum_{\{i,j\} \in E} |e^{-ie_{ij}} - 1|^2 \leq |e|_{\mathbb{T}}^2,$$

and so combining this with (5.5.6) gives

$$\sum_{\{i,j\} \in E} |e^{i(\gamma_j - \gamma_i)} - 1|^2 \leq 2\left(\frac{\pi^2}{4} \frac{C'}{\lambda_2} + 1\right) |e|_{\mathbb{T}}^2. \quad (5.5.7)$$

At this point, take $\alpha := \frac{1}{\text{vol}(G)} \sum_{i \in V} \deg(i) e^{i\gamma_i}$, define v entrywise by $v_i := e^{i\gamma_i}$, and set $w = v - \alpha 1$. Then

$$1^* D w = \sum_{i \in V} \deg(i) (e^{i\gamma_i} - \alpha) = 0,$$

i.e., $D^{1/2}w$ is orthogonal to $D^{1/2}1$. Also, since $(D - A)1 = 0$, we have

$$L(D^{1/2}1) = (I - D^{-1/2}AD^{-1/2})D^{1/2}1 = D^{-1/2}(D - A)D^{-1/2}D^{1/2}1 = 0,$$

and so $D^{1/2}w$ is orthogonal to a first eigenvector of L , considering L is positive semidefinite. Thus,

$$\frac{(D^{1/2}w)^* L(D^{1/2}w)}{w^* D w} \geq \min_{\substack{y \in \mathbb{C}^{|V|} \\ y \perp D^{1/2}1}} \frac{y^* L y}{y^* y} = \lambda_2.$$

Rearranging then gives

$$\lambda_2 w^* D w \leq (D^{1/2}w)^* L(D^{1/2}w) = (v - \alpha 1)^* (D - A)(v - \alpha 1) = v^* (D - A)v.$$

Continuing, we apply the definitions of D and A to get

$$\lambda_2 w^* D w \leq 2 |E| - \sum_{i \in V} \sum_{j \in V} e^{-i\gamma_j} A[i, j] e^{i\gamma_i} = \sum_{\{i, j\} \in E} \left(2 - e^{i(\gamma_i - \gamma_j)} - e^{i(\gamma_j - \gamma_i)} \right) = \sum_{\{i, j\} \in E} |e^{i\gamma_j} - e^{i\gamma_i}|^2.$$

Next, we factor $e^{i\gamma_i}$ from the inside of each term and apply (5.5.7) to get

$$\lambda_2 \sum_{i \in V} \deg(i) |e^{i\gamma_i} - \alpha|^2 = \lambda_2 w^* D w \leq \sum_{\{i, j\} \in E} |e^{i(\gamma_j - \gamma_i)} - 1|^2 \leq 2 \left(\frac{\pi^2}{4} \frac{C'}{\lambda_2} + 1 \right) |e|_{\mathbb{T}}^2. \quad (5.5.8)$$

From here, we proceed in two cases. First, when $\alpha \neq 0$, we may take $\theta := \arg(\alpha)$.

Then the left-hand inequality of (5.5.4) and Lemma 5.5.6 give

$$\sum_{i \in V} |\gamma_i - \theta|_{\mathbb{T}}^2 \leq \frac{\pi^2}{4} \sum_{i \in V} |e^{i(\gamma_i - \theta)} - 1|^2 = \frac{\pi^2}{4} \sum_{i \in V} |e^{i\gamma_i} - \alpha|^2 \leq \pi^2 \sum_{i \in V} |e^{i\gamma_i} - \alpha|^2 \leq \pi^2 \sum_{i \in V} \deg(i) |e^{i\gamma_i} - \alpha|^2 \quad (5.5.9)$$

where the last inequality uses the fact that $\deg(i) \geq 1$ for every $i \in V$, i.e., the graph G has no isolated vertex since G is connected, which follows from the fact that $\lambda_2 > 0$.

In the case where $\alpha = 0$, we may arbitrarily take $\theta = 0$. Then similar analysis yields

$$\sum_{i \in V} |\gamma_i - \theta|_{\mathbb{T}}^2 \leq \frac{\pi^2}{4} \sum_{i \in V} |e^{i\gamma_i} - 1|^2 \leq \pi^2 |V| \leq \pi^2 \sum_{i \in V} \deg(i) = \pi^2 \sum_{i \in V} \deg(i) |e^{i\gamma_i} - \alpha|^2,$$

where in this case, the second inequality applies the triangle inequality to each term instead of Lemma 5.5.6. Note that both cases produce the same bound on $\sum_{i \in V} |\gamma_i - \theta|_{\mathbb{T}}^2$, which we now conclude by combining (5.5.8) and (5.5.9):

$$\sum_{i \in V} |\gamma_i - \theta|_{\mathbb{T}}^2 \leq \frac{2\pi^2}{\lambda_2} \left(\frac{\pi^2}{4} \frac{C'}{\lambda_2} + 1 \right) |e|_{\mathbb{T}}^2 \leq \frac{2\pi^2}{\lambda_2} \left(\frac{\pi^2}{4} \frac{C'}{\lambda_2} + 1 \right) \cdot \frac{\pi^2 |\varepsilon|^2}{P^2}.$$

The last inequality follows from Lemma 5.3.3, taking $z = \overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle + \varepsilon_{ij}$ and

$$\varepsilon = -\varepsilon_{ij}. \quad \square$$

5.5.3 Projective uniformity

This section is motivated by Theorem 5.5.3 of the previous section, which exhibits significant dependence on the size of P . Here, we show how Algorithm 5.3.2 ensures that P will not too small, and our guarantee will be in terms of the following noise-robust version of projective uniformity:

Definition 5.5.7 (Projective uniformity with noise). Consider a graph $G = (V, E)$ and $M \times |V|$ matrix Φ , and for some proportion $\alpha \in (0, 1)$, signal $x \in \mathbb{C}^M$ and noise $\varepsilon = \{\varepsilon_{ij}\}_{\{i,j\} \in E}$, let $\mathcal{J}(\alpha, x, \varepsilon)$ denote the set of vertices that remain after applying Algorithm 5.3.2. Then the *projective uniformity with noise* of Φ is

$$\text{PUN}(\Phi; \alpha, \varepsilon) = \min_{\substack{x \in \mathbb{C}^M \\ |x|=1}} \min_{\{i,j\} \in \mathcal{J}(\alpha, x, \varepsilon)} \left| \overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle + \varepsilon_{ij} \right|.$$

Theorem 5.5.8. Consider a graph $G = (V, E)$ with $|V| \sim CM \log M$ for some constant C , and draw the entries of an $M \times |V|$ matrix Φ independently from $\mathbb{CN}(0, \frac{1}{M})$. Then for each proportion $\alpha < 1 - \frac{1}{2C}$, there exists a constant $C' > 0$ such that, with overwhelming probability,

$$\max \left\{ \|\varepsilon\|_2, \sqrt{M} \text{PUN}(\Phi; \alpha, \varepsilon) \right\} \geq \frac{C'}{\sqrt{M}}$$

for every noise vector $\varepsilon = \{\varepsilon_{ij}\}_{\{i,j\} \in E}$.

To prove this theorem, we apply the following lemma, which follows from concentration-of-measure arguments that we provide later:

Lemma 5.5.9. Take $n \sim CM \log M$ for some constant C , and draw the entries of an $M \times n$ matrix Φ independently from $\mathbb{CN}(0, \frac{1}{M})$. Then for each proportion $\alpha < 1 - \frac{1}{2C}$, there exists a constant $C' > 0$ such that

$$\text{PU}(\Phi; \alpha) \geq \frac{C'}{M}$$

with overwhelming probability.

Proof of Theorem 5.5.8. Denote $n := |V|$. Note that if

$$|\varepsilon|_2 > \sqrt{\frac{1-\alpha}{2}n} \cdot \frac{1}{2}\text{PU}(\Phi; \alpha + \frac{1+\alpha}{2}), \quad (5.5.10)$$

then by Lemma 5.5.9, we have $|\varepsilon|_2 \geq \frac{C'}{\sqrt{M}}$ with overwhelming probability for some constant $C' > 0$. It remains to consider the case where (5.5.10) does not hold. To this end, define

$$\mathcal{J}_1(\alpha, x, \varepsilon) := \operatorname{argmax}_{\substack{\mathcal{J} \subseteq \{1, \dots, n\} \\ |\mathcal{J}| \geq \alpha n}} \min_{\{i, j\} \in \mathcal{J}} \left| \overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle + \varepsilon_{ij} \right|.$$

Recalling the definition of $\mathcal{J}(\alpha, x, \varepsilon)$ in Definition 5.5.7, we claim that $\mathcal{J}(\alpha, x, \varepsilon) \subseteq \mathcal{J}_1(\alpha, x, \varepsilon)$. To see this, label each edge $\{i, j\} \in E$ with $\left| \overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle + \varepsilon_{ij} \right|$. Then by definition, $V \setminus \mathcal{J}_1(\alpha, x, \varepsilon)$ neighbors more of the smallest edges than any other collection of $n - \lceil \alpha n \rceil = \lfloor (1 - \alpha)n \rfloor$ vertices. By comparison, Algorithm 5.3.2 effectively deletes the smallest edge $\{i, j\}$ by deleting both of its incident vertices i and j , one of which must be in $\mathcal{J}_1(\alpha, x, \varepsilon)$. The smallest edge in the remaining graph is guaranteed to not touch either i or j , but rather some $k \in \mathcal{J}_1(\alpha, x, \varepsilon)$, provided $\mathcal{J}_1(\alpha, x, \varepsilon)$ has not yet been completely removed from the graph. After $\lfloor (1 - \alpha)n \rfloor$ iterations, then by the pigeonhole principle, all vertices in $V \setminus \mathcal{J}_1(\alpha, x, \varepsilon)$ will be removed, meaning $\mathcal{J}(\alpha, x, \varepsilon) \subseteq \mathcal{J}_1(\alpha, x, \varepsilon)$, as claimed. This implies that

$$\min_{\{i, j\} \in \mathcal{J}(\alpha, x, \varepsilon)} \left| \overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle + \varepsilon_{ij} \right| \geq \min_{\{i, j\} \in \mathcal{J}_1(\alpha, x, \varepsilon)} \left| \overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle + \varepsilon_{ij} \right|,$$

and so minimizing over all unit vectors x gives

$$\text{PUN}(\Phi; \alpha, \varepsilon) \geq \min_{\substack{x \in \mathbb{C}^M \\ |x|=1}} \max_{\substack{\mathcal{J} \subseteq \{1, \dots, n\} \\ |\mathcal{J}| \geq \alpha n}} \min_{\{i, j\} \in \mathcal{J}} \left| \overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle + \varepsilon_{ij} \right|. \quad (5.5.11)$$

Next, define

$$\mathcal{J}_2(\alpha, x, \varepsilon) := \operatorname{argmax}_{\substack{\mathcal{J} \subseteq \{1, \dots, n\} \\ |\mathcal{J}| \geq (\alpha + \frac{1-\alpha}{2})n}} \min_{\{i, j\} \in \mathcal{J}} \left| \overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle \right|.$$

Since in this case (5.5.10) does not hold, there are at most $\frac{1-\alpha}{2}n$ edges $\{i, j\} \in E$ with $|\varepsilon_{ij}| \geq \frac{1}{2}\text{PU}(\Phi; \alpha + \frac{1+\alpha}{2})$. Some of these edges are induced by $\mathcal{J}_2(\alpha, x, \varepsilon)$; as such, delete at most $\frac{1-\alpha}{2}n$ vertices from $\mathcal{J}_2(\alpha, x, \varepsilon)$ which are incident to all of these edges, and denote the remaining vertices by $\mathcal{J}_3(\alpha, x, \varepsilon)$. By construction, we have $|\mathcal{J}_3(\alpha, x, \varepsilon)| \geq \alpha n$, and also the edges $\{i, j\}$ induced by $\mathcal{J}_3(\alpha, x, \varepsilon)$ satisfy $|\varepsilon_{ij}| < \frac{1}{2}\text{PU}(\Phi; \alpha + \frac{1+\alpha}{2})$. These facts combined with the triangle inequality then give

$$\begin{aligned} \max_{\substack{\mathcal{J} \subseteq \{1, \dots, n\} \\ |\mathcal{J}| \geq \alpha n}} \min_{\{i, j\} \in \mathcal{J}} \left| \overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle + \varepsilon_{ij} \right| &\geq \min_{\{i, j\} \in \mathcal{J}_3(\alpha, x, \varepsilon)} \left| \overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle + \varepsilon_{ij} \right| \\ &\geq \min_{\{i, j\} \in \mathcal{J}_3(\alpha, x, \varepsilon)} \left(\left| \overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle \right| - |\varepsilon_{ij}| \right) \\ &\geq \min_{\{i, j\} \in \mathcal{J}_3(\alpha, x, \varepsilon)} \left| \overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle \right| - \frac{1}{2}\text{PU}(\Phi; \alpha + \frac{1+\alpha}{2}). \end{aligned} \tag{5.5.12}$$

Continuing, we use the fact that $\mathcal{J}_3(\alpha, x, \varepsilon) \subseteq \mathcal{J}_2(\alpha, x, \varepsilon)$ to get

$$\begin{aligned} \min_{\{i, j\} \in \mathcal{J}_3(\alpha, x, \varepsilon)} \left| \overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle \right| &\geq \min_{\{i, j\} \in \mathcal{J}_2(\alpha, x, \varepsilon)} \left| \overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle \right| \\ &\geq \min_{\substack{x \in \mathbb{C}^M \\ |x|=1}} \max_{\substack{\mathcal{J} \subseteq \{1, \dots, n\} \\ |\mathcal{J}| \geq (\alpha + \frac{1-\alpha}{2})n}} \min_{\{i, j\} \in \mathcal{J}} \left| \overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle \right| \\ &\geq \text{PU}(\Phi; \alpha + \frac{1+\alpha}{2}). \end{aligned} \tag{5.5.13}$$

Combining (5.5.11), (5.5.12) and (5.5.13) then gives $\text{PUN}(\Phi; \alpha, \varepsilon) \geq \frac{1}{2}\text{PU}(\Phi; \alpha + \frac{1+\alpha}{2})$, and so the result follows directly from Lemma 5.5.9. $\square$

We conclude this section with the rather technical proof of Lemma 5.5.9:

Proof of Lemma 5.5.9. We seek a bound on the probability that $\text{PU}(\Phi; \alpha) < \frac{C'}{M}$. To

do so, we will cover this event with smaller failure events $\mathcal{E}_k$ which correspond to members of a δ -net. First, define

$$G_\delta(v) := \{\varphi \in \mathbb{C}^M : |\langle v, \varphi \rangle| \geq 3\delta \text{ and } |\varphi| \leq 2\}.$$

Then for each member v of a given δ -net $\mathcal{N}_\delta$ of the unit sphere in $\mathbb{C}^M$, define the failure event

$$\mathcal{E}_v := \left\{ \text{less than } \alpha n \text{ columns of } \Phi \text{ lie in } G_\delta(v) \right\}.$$

We claim that $\{\text{PU}(\Phi; \alpha) < \delta^2\} \subseteq \bigcup_{v \in \mathcal{N}_\delta} \mathcal{E}_v$. To see this, take an outcome $\omega \in (\bigcup_{v \in \mathcal{N}_\delta} \mathcal{E}_v)^c$. By the definition of $\mathcal{N}_\delta$, we have that for every unit vector $x \in \mathbb{C}^M$, there exists $v \in \mathcal{N}_\delta$ such that $|v - x| \leq \delta$. This gives a mapping $x \mapsto v_x$, and we let $\mathcal{I}_x$ denote the indices of columns of Φ which lie in $G_\delta(v_x)$. Since $\omega \in \bigcap_{v \in \mathcal{N}_\delta} \mathcal{E}_v^c$, we have $|\mathcal{I}_x| \geq \alpha n$ for every x . Moreover, by the triangle inequality, Cauchy-Schwarz inequality, and the definition of G_δ , every $i \in \mathcal{I}_x$ satisfies

$$|\langle x, \varphi_i \rangle| \geq |\langle v_x, \varphi_i \rangle| - |\langle v_x - x, \varphi_i \rangle| \geq |\langle v_x, \varphi_i \rangle| - |v_x - x| |\varphi_i| \geq 3\delta - 2|v_x - x| \geq \delta.$$

Applying this inequality then gives

$$\text{PU}(\Phi; \alpha) = \min_{\substack{x \in \mathbb{C}^M \\ |x|=1}} \max_{\substack{\mathcal{I} \subseteq \{1, \dots, n\} \\ |\mathcal{I}| \geq \alpha n}} \min_{i \in \mathcal{I}} |\langle x, \varphi_i \rangle|^2 \geq \min_{\substack{x \in \mathbb{C}^M \\ |x|=1}} \min_{i \in \mathcal{I}_x} |\langle x, \varphi_i \rangle|^2 \geq \delta^2.$$

Thus $\omega \in \{\text{PU}(\Phi; \alpha) < \delta^2\}^c$, thereby proving our claim. Continuing, the union bound gives

$$\Pr\left(\text{PU}(\Phi; \alpha) < \delta^2\right) \leq \sum_{v \in \mathcal{N}_\delta} \Pr(\mathcal{E}_v) = |\mathcal{N}_\delta| \cdot \Pr(\mathcal{E}_v),$$

where the equality follows from the fact that the columns of Φ are independent with rotationally symmetric probability distributions. It therefore suffices to bound both $|\mathcal{N}_\delta|$ and $\Pr(\mathcal{E}_v)$.

To bound $|\mathcal{N}_\delta|$, we follow a standard argument, found in the proof of Lemma 5.2 in [Ver10]. Let $\mathcal{N}_\delta$ be a maximal δ -packing of points on the unit sphere in $\mathbb{C}^M$. Since the packing is maximal, it follows that $\mathcal{N}_\delta$ is a δ -net. To count these points, map them into $\mathbb{R}^{2M}$ according to $f: v \mapsto (\operatorname{Re} v, \operatorname{Im} v)$. Consider the open balls of radius $\frac{\delta}{2}$ centered at each $f(v)$. Note that the (disjoint) union of these balls is contained in the ball of radius $1 + \frac{\delta}{2}$ centered at the origin. Thus, a volume comparison gives

$$|\mathcal{N}_\delta| \cdot C\left(\frac{\delta}{2}\right)^{2M} = \operatorname{Vol}\left(\bigsqcup_{v \in \mathcal{N}_\delta} B(f(v), \frac{\delta}{2})\right) \leq \operatorname{Vol}\left(B(0, 1 + \frac{\delta}{2})\right) = C\left(1 + \frac{\delta}{2}\right)^{2M},$$

thereby implying $|\mathcal{N}_\delta| \leq \left(\frac{2}{\delta} + 1\right)^{2M}$.

To bound $\Pr(\mathcal{E}_v)$, note that

$$\Pr(\mathcal{E}_v) = \Pr\left(\sum_{i=1}^n 1_{\{\varphi_i \notin G_\delta(v)\}} \geq (1 - \alpha)n\right). \quad (5.5.14)$$

As such, we first consider the success probability of these Bernoulli random variables:

$$\Pr(\varphi_i \notin G_\delta(v)) \leq \Pr(|\langle v, \varphi_i \rangle| < 3\delta) + \Pr(|\varphi_i| > 2). \quad (5.5.15)$$

By the rotational symmetry of the distribution of φ_i , we may take v to be the first identity basis element e_1 without changing the probability. Next, since $\varphi_i = \sum_{m=1}^M (a_m + ib_m)e_m$ with the a_m 's and b_m 's independent with distribution $\mathcal{N}(0, \frac{1}{2M})$, we have

$$\Pr(|\langle v, \varphi_i \rangle| < 3\delta) = \Pr(\sqrt{a_1^2 + b_1^2} < 3\delta) \leq \Pr(\min\{a_1^2, b_1^2\} < \frac{9}{2}\delta^2) \leq 2\Pr(a_1^2 < \frac{9}{2}\delta^2),$$

where the last step is by the union bound. Since a_1 's density function is $\leq \sqrt{\frac{M}{\pi}}$, it follows that

$$\Pr(|\langle v, \varphi_i \rangle| < 3\delta) \leq 2\Pr(|a_1| < \frac{3}{\sqrt{2}}\delta) \leq 12\delta\sqrt{\frac{M}{2\pi}}. \quad (5.5.16)$$

For the other term in (5.5.15), note that $2M |\varphi_i|^2$ is a sum of independent standard Gaussian random variables. Applying Lemma 1 of [LM00] then gives that for every $t > 0$,

$$\Pr\left(2M |\varphi_i|^2 \geq \sqrt{8Mt} + 2t + 2M\right) \leq e^{-t}.$$

Thus, taking $t = \frac{M}{2}$ gives

$$\Pr\left(|\varphi_i| > 2\right) = \Pr\left(2M |\varphi_i|^2 > 8M\right) \leq \Pr\left(2M |\varphi_i|^2 \geq 5M\right) \leq e^{-M/2}. \quad (5.5.17)$$

Substituting (5.5.16) and (5.5.16) into (5.5.15) then gives

$$\Pr\left(\varphi_i \notin G_\delta(v)\right) \leq 12\delta\sqrt{\frac{M}{2\pi}} + e^{-M/2}. \quad (5.5.18)$$

Now, to bound (5.5.14), we will apply Hoeffding's inequality [Hoe63], which says that the tail probability of a sum of independent Bernoulli random variables X_i , each with success probability p , has the following bound:

$$\Pr\left(\sum_{i=1}^n X_i \geq n(p+t)\right) \leq e^{-2nt}.$$

Also, note that replacing p in the left-hand side above with some $p' \geq p$ will not increase the probability. As such, taking p' to be the right-hand side of (5.5.18) and $t = 1 - \alpha - p'$, we have

$$\Pr(\mathcal{E}_v) = \Pr\left(\sum_{i=1}^n 1_{\{\varphi_i \notin G_\delta(v)\}} \geq (1-\alpha)n\right) \leq \exp\left(-2n\left((1-\alpha) - \left(12\delta\sqrt{\frac{M}{2\pi}} + e^{-M/2}\right)\right)\right).$$

Now that we have bounds on both $|\mathcal{N}_\delta|$ and $\Pr(\mathcal{E}_v)$, we note that we specifically wish to show that

$$\Pr\left(\text{PU}(\Phi; \alpha) < \frac{C'}{M}\right) < c_0 e^{-c_1 M}$$

for some constants $c_0, c_1 > 0$. Considering the above analysis and taking logarithms,

it suffices to have

$$\log |\mathcal{N}_\delta| + \log \Pr(\mathcal{E}_v) \leq 2M \log\left(\frac{2}{\delta} + 1\right) - 2n(1 - \alpha) + 2n \left(12\delta \sqrt{\frac{M}{2\pi}} + e^{-M/2}\right) < \log c_0 - c_1 M$$

with $\delta^2 = \frac{C'}{M}$. Since $\alpha < 1 - \frac{1}{2C}$ by assumption, picking $C' < \frac{\pi}{72} \left(1 - \frac{1}{2C} - \alpha\right)^2$ will make $2n(1 - \alpha)$ the dominant term in the above inequality, thereby proving the result. $\square$

5.5.4 Removing large vertices

In this section, we prove how well we can remove the vertices with the largest noisy intensity measurements. We start with a lemma:

Lemma 5.5.10. *Pick $C > 0$ and $n \geq e^{C/8} M \log M$. Draw the entries of an $M \times n$ matrix $\Phi = \{\varphi_i\}_{i=1}^n$ independently from $\mathbb{CN}(0, \frac{1}{M})$. Taking $\beta := 3e^{-C/8}$, then with overwhelming probability,*

$$\#\left\{i : |\langle x, \varphi_i \rangle|^2 > \frac{C}{M}\right\} < \beta n \quad (5.5.19)$$

for every unit norm $x \in \mathbb{C}^M$.

Proof. Take a δ -net $\mathcal{N}_\delta$ of the unit sphere in $\mathbb{C}^M$. Then for every unit vector x , there is a $v_x \in \mathcal{N}_\delta$ such that

$$|\langle x, \varphi_i \rangle| \leq |\langle v_x, \varphi_i \rangle| + |\langle x - v_x, \varphi_i \rangle| \leq |\langle v_x, \varphi_i \rangle| + |x - v_x| \leq |\langle v_x, \varphi_i \rangle| + \delta.$$

Taking $\delta = \frac{1}{2} \sqrt{\frac{C}{M}}$, we then have that $|\langle x, \varphi_i \rangle|^2 > \frac{C}{M}$ implies $|\langle v_x, \varphi_i \rangle|^2 > \frac{C}{4M}$. Recall that we wish to bound the probability of the event $\mathcal{E}$ that (5.5.19) is violated for some unit vector x . To this end, the above implication allows us to focus on a finite set of points:

$$\Pr(\mathcal{E}) \leq \Pr\left(\exists v \in \mathcal{N}_\delta \text{ s.t. } \sum_{i=1}^n 1_{\{|\langle v, \varphi_i \rangle|^2 > \frac{C}{4M}\}} \geq \beta n\right).$$

As discussed in the proof of Lemma 5.5.9, we may take $|\mathcal{N}_\delta| \leq (\frac{2}{\delta} + 1)^{2M}$, and so the union bound and the symmetric distribution of Φ both give

$$\Pr(\mathcal{E}) \leq \left(4\sqrt{\frac{M}{C}} + 1\right)^{2M} \Pr\left(\sum_{i=1}^n 1_{\{|\langle v, \varphi_i \rangle|^2 > \frac{C}{4M}\}} \geq \beta n\right). \quad (5.5.20)$$

Similar to the proof of Lemma 5.5.9, we will bound the probability on the right-hand side using Hoeffding's inequality. First, we note that the symmetric distribution of $\varphi_i = \sum_{m=1}^M (a_m + ib_m)e_m$ gives that

$$\Pr\left(|\langle v, \varphi_i \rangle|^2 > \frac{C}{4M}\right) = \Pr\left(|\langle e_1, \varphi_i \rangle|^2 > \frac{C}{4M}\right) = \Pr\left(a_1^2 + b_1^2 > \frac{C}{4M}\right) \leq \Pr\left(a_1^2 > \frac{C}{8M}\right) + \Pr\left(b_1^2 > \frac{C}{8M}\right),$$

where the last step is by the union bound. We continue, using the fact that a_1 and b_1 are both distributed as $\mathcal{N}(0, \frac{1}{2M})$:

$$\Pr\left(|\langle v, \varphi_i \rangle|^2 > \frac{C}{4M}\right) \leq 2\Pr\left(a_1^2 > \frac{C}{8M}\right) = 2\Pr\left(z^2 > \frac{C}{4}\right) \leq 2e^{-C/8},$$

where z is a standard Gaussian random variable. With this, we now apply Hoeffding's inequality to (5.5.20):

$$\Pr(\mathcal{E}) \leq \left(4\sqrt{\frac{M}{C}} + 1\right)^{2M} \exp\left(-2n(\beta - 2e^{-C/8})\right) = \exp\left(2M \log\left(4\sqrt{\frac{M}{C}} + 1\right) - 2ne^{-C/8}\right).$$

Since $n \geq e^{C/8} M \log M$, then $2ne^{-C/8}$ is the dominant term above, thereby proving the result. $\square$

Theorem 5.5.11. *Pick $C > 0$ and $n \geq e^{C/8} M \log M$. Draw the entries of an $M \times n$ matrix $\Phi = \{\varphi_i\}_{i=1}^n$ independently from $\mathbb{CN}(0, \frac{1}{M})$. Given a vector $x \in \mathbb{C}^M$, consider*

$$z_i := |\langle x, \varphi_i \rangle|^2 + \nu_i,$$

where $\nu = \{\nu_i\}_{i=1}^n$ satisfies $\frac{|\nu|}{|x|^2} \leq \frac{C'}{\sqrt{M}}$. Taking $\beta := 3e^{-C/8}$, we have

$$\#\left\{i : z_i > \frac{2C}{M} |x|^2\right\} < 2\beta n \quad (5.5.21)$$

with overwhelming probability.

Proof. In counting large z_i 's, we identify which come from large or small inner products $|\langle x, \varphi_i \rangle|^2$. First,

$$\left\{i : z_i > \frac{2C}{M} |x|^2\right\} \cap \left\{i : |\langle x, \varphi_i \rangle|^2 > \frac{C}{M} |x|^2\right\}$$

is of size $< \beta n$ with overwhelming probability by Lemma 5.5.10. The rest of the large z_i 's have indices in

$$\mathcal{K} := \left\{i : z_i > \frac{2C}{M} |x|^2\right\} \cap \left\{i : |\langle x, \varphi_i \rangle|^2 \leq \frac{C}{M} |x|^2\right\}.$$

To count these, note that $\nu_i = z_i - |\langle x, \varphi_i \rangle|^2 \geq \frac{C}{M} |x|^2$, and so

$$|\mathcal{K}| \frac{C^2}{M^2} |x|^4 \leq \sum_{i \in \mathcal{K}} |\nu_i|^2 \leq |\nu|^2 \leq \frac{(C')^2}{M} |x|^4.$$

Rearranging then reveals that $|\mathcal{K}| = \mathcal{O}(M)$, meaning $\mathcal{K}$ has fewer than $\beta n \geq 3M \log M$ members when M is sufficiently large. $\square$

5.5.5 Main result

This section proves the main result of the paper, which we restate here:

Theorem 5.5.12. *Pick $N \sim CM \log M$ with C sufficiently large, and take $\{\varphi_\ell\}_{\ell=1}^N = \Phi_V \cup \Phi_E$ defined in the measurement design of Section 5.3. Then there exist constants $C', K > 0$ such that the following guarantee holds for all $x \in \mathbb{C}^M$ with overwhelming*

probability: Consider measurements of the form

$$z_\ell := |\langle x, \varphi_\ell \rangle|^2 + \nu_\ell.$$

If the noise-to-signal ratio satisfies $\text{NSR} := \frac{|\nu|}{|x|^2} \leq \frac{C'}{\sqrt{M}}$, then the phase retrieval procedure of Section 5.3 produces an estimate $\tilde{x}$ from $\{z_\ell\}_{\ell=1}^N$ with squared relative error

$$\frac{|\tilde{x} - e^{i\theta}x|^2}{|x|^2} \leq K \sqrt{\frac{M}{\log M}} \text{NSR}$$

for some phase $\theta \in [0, 2\pi)$.

Proof. We will prove the result by considering the steps of our phase retrieval process in reverse order. In the last step, we have the following estimates of $\langle x, \varphi_i \rangle$ for every vertex $i \in V'' \subseteq V$ which survives our graph-pruning and large-vertex-removing processes:

$$y_i := e^{i\theta} \langle x, \varphi_i \rangle + \delta_i = (e^{i\theta} \Phi_{V''}^* x + \delta)_i.$$

Here, θ is a global phase which is calculated in the proof of Theorem 5.5.3. From these estimates, we reconstruct by finding the least-squares estimate of x :

$$\tilde{x} := (\Phi_{V''} \Phi_{V''}^*)^{-1} \Phi_{V''} y = e^{i\theta} x + (\Phi_{V''} \Phi_{V''}^*)^{-1} \Phi_{V''} \delta.$$

As such, we have the following bound on the reconstruction error:

$$|\tilde{x} - e^{i\theta}x| = |(\Phi_{V''} \Phi_{V''}^*)^{-1} \Phi_{V''} \delta| \leq |(\Phi_{V''} \Phi_{V''}^*)^{-1} \Phi_{V''}|_2 |\delta| = \frac{|\delta|}{\sigma_{\min}(\Phi_{V''}^*)}, \quad (5.5.22)$$

where the last equality holds with probability 1, specifically, in the event that $\sigma_{\min}(\Phi_{V''}^*) > 0$.

To continue this bound, recall that the large-vertex-removing process ensured $|V''| = \kappa n$. We claim there exists a constant $c_0 > 0$ such that with overwhelming

probability, $\sigma_{\min}(\Phi_{V''}^*) \geq c_0 \sqrt{\log M}$ for every $V'' \subseteq V$ with $|V''| = \kappa n$. To prove this, we leverage the complex version of Corollary 5.35 in [Ver10], which gives that for a fixed $V'' \subseteq V$, there exist $c, c' > 0$ such that

$$\Pr\left(\sigma_{\min}(\Phi_{V''}^*) < \frac{1}{\sqrt{M}}(\sqrt{\kappa n} - \sqrt{M} - \sqrt{tn})\right) \leq ce^{-c'tn}.$$

Performing a union bound over the $\binom{n}{\kappa n} = \binom{n}{(1-\kappa)n} \leq (\frac{e}{1-\kappa})^{(1-\kappa)n}$ choices for V'' then gives

$$\Pr\left(\exists V'' \subseteq V, |V''| = \kappa n \text{ s.t. } \sigma_{\min}(\Phi_{V''}^*) < \frac{1}{\sqrt{M}}(\sqrt{\kappa n} - \sqrt{M} - \sqrt{tn})\right) \leq c \exp\left(c'n\left(-\frac{t}{2} + (1-\kappa) \log\left(\frac{e}{1-\kappa}\right)\right)\right)$$

Provided $\frac{\kappa}{2} > (1-\kappa) \log(\frac{e}{1-\kappa})$, which occurs whenever $\kappa \geq 0.86$, then taking $t := \frac{\kappa}{2} + (1-\kappa) \log(\frac{e}{1-\kappa})$ will ensure both $t < \kappa$ and $\frac{t}{2} > (1-\kappa) \log(\frac{e}{1-\kappa})$. Therefore, this choice for t coupled with the fact that $n = CM \log M$ proves the claim. Continuing (5.5.22) then gives

$$|\tilde{x} - e^{i\theta}x| \leq \frac{|\delta|}{c_0 \sqrt{\log M}}. \quad (5.5.23)$$

Next, we wish to bound $|\delta|$. By definition, we have

$$\delta_i = y_i - e^{i\theta} \langle x, \varphi_i \rangle = \sqrt{z_i} e^{i \arg(u_i)} - e^{i\theta} \langle x, \varphi_i \rangle.$$

Note that the above square root operates under the assumption that $z_i \geq 0$ for each $i \in V''$, which is ensured when we prune for reliability. Denote $\xi_i := \sqrt{z_i} - |\langle x, \varphi_i \rangle|$. Then by the triangle inequality, we have

$$|\delta_i| = \left| \sqrt{z_i} e^{i \arg(u_i)} - \sqrt{z_i} e^{i(\theta + \arg(\langle x, \varphi_i \rangle))} + \xi_i e^{i(\theta + \arg(\langle x, \varphi_i \rangle))} \right| \leq \sqrt{z_i} \left| e^{i \arg(u_i)} - e^{i(\theta + \arg(\langle x, \varphi_i \rangle))} \right| + |\xi_i|.$$

Next, factoring $e^{i(\theta + \arg(\langle x, \varphi_i \rangle))}$ and applying (5.5.4) yields

$$|\delta_i| \leq \sqrt{z_i} |\arg(u_i) - \arg(\langle x, \varphi_i \rangle) - \theta|_{\mathbb{T}} + |\xi_i|.$$

For any $a, b \geq 0$, then since $0 \leq (a - b)^2 = a^2 - 2ab + b^2$, we have $(a + b)^2 = a^2 + 2ab + b^2 \leq 2(a^2 + b^2)$. Applying this inequality to the right-hand side above then gives

$$|\delta_i|^2 \leq 2z_i |\arg(u_i) - \arg(\langle x, \varphi_i \rangle) - \theta|_{\mathbb{T}}^2 + 2\xi_i^2.$$

Similarly, for any $a, b \geq 0$, then since $ab \geq \min\{a^2, b^2\}$, we have $(a - b)^2 = a^2 - 2ab + b^2 \leq |a^2 - b^2|$. Applying this to $\xi_i^2 = (\sqrt{z_i} - |\langle x, \varphi_i \rangle|)^2$ then gives $\xi_i^2 \leq |\nu_i|$. Also by Theorem 5.5.11, our large-vertex-removing process ensures that $z_i \leq \frac{2c_1}{M} |x|^2$ for every $i \in V''$. Combined, these facts imply

$$|\delta|^2 \leq \frac{4c_1}{M} |x|^2 \sum_{i \in V''} |\arg(u_i) - \arg(\langle x, \varphi_i \rangle) - \theta|_{\mathbb{T}}^2 + 2|\nu_V|_1. \quad (5.5.24)$$

Applying Theorem 5.5.3, Definition 5.5.7 and Theorem 5.5.8 further gives

$$\sum_{i \in V''} |\arg(u_i) - \arg(\langle x, \varphi_i \rangle) - \theta|_{\mathbb{T}}^2 \leq \frac{c_2 |\varepsilon|^2}{\lambda_2^2 P^2} \leq \frac{c_2 |\varepsilon|^2}{\lambda_2^2 |x|^4 \text{PUN}(\Phi_V; \alpha, \frac{\varepsilon}{|x|^2})^2} \leq \frac{c_2 M^2 |\varepsilon|^2}{c_3^2 \lambda_2^2 |x|^4}. \quad (5.5.25)$$

Recalling the definition of ε , we have

$$\varepsilon_{ij} = \left(\frac{1}{3} \sum_{k=0}^2 \omega^k \left(|\langle x, \varphi_i + \omega^k \varphi_j \rangle|^2 + \nu_{ijk} \right) \right) - \overline{\langle x, \varphi_i \rangle} \langle x, \varphi_j \rangle = \frac{1}{3} \sum_{k=0}^2 \omega^k \nu_{ijk},$$

and so, letting E' denote the edges which remained after pruning for connectivity, the Cauchy-Schwarz inequality gives

$$|\varepsilon|^2 = \sum_{\{i,j\} \in E'} |\varepsilon_{ij}|^2 = \sum_{\{i,j\} \in E'} \left| \frac{1}{3} \sum_{k=0}^2 \omega^k \nu_{ijk} \right|^2 \leq \frac{1}{9} \sum_{\{i,j\} \in E'} \left(\sum_{k=0}^2 |\omega^{-k}|^2 \right) \left(\sum_{k=0}^2 |\nu_{ijk}|^2 \right) \leq \frac{1}{3} |\nu|^2. \quad (5.5.26)$$

Finally, we combine (5.5.23), (5.5.24), (5.5.25) and (5.5.26), along with $|\nu_V|_1 \leq \sqrt{n} |\nu_V| \leq \sqrt{2CM \log M} |\nu|$:

$$\frac{|\tilde{x} - e^{i\theta} x|^2}{|x|^2} \leq \frac{4c_1 c_2}{3c_0^2 c_3^2 \lambda_2^2 \log M} \frac{M}{|x|^4} |\nu|^2 + \frac{2\sqrt{2C}}{c_0^2} \sqrt{\frac{M}{\log M}} \frac{|\nu|}{|x|^2} \leq K \sqrt{\frac{M}{\log M}} \frac{|\nu|}{|x|^2},$$

for some constant K . □

Chapter 6

Forbidden induced subgraphs of double-split graphs

6.1 Introduction

A key ingredient in the proof of the strong perfect graph theorem by Chudnovsky, Robertson, Seymour, and Thomas [CRST06] is a decomposition theorem for all perfect graphs. This decomposition theorem states that all perfect graphs either belong to one of five basic classes or admit one of several decompositions. The five basic classes are bipartite graphs, complements of bipartite graphs, line graphs of bipartite graphs, complements of line graphs of bipartite graphs, and double-split graphs. The first four classes are closed under taking induced subgraphs and have known characterizations in terms of minimal forbidden induced subgraphs. Indeed, a forbidden induced subgraph characterization is known for the union of these four classes [ZZ05]. However, double-split graphs are not closed under taking induced subgraphs, and hence do not have such a characterization.

In this paper, we consider the downward closure of double-split graphs under induced subgraphs (that is, double-split graphs and all of their induced subgraphs)

and we characterize this class in terms of minimal forbidden induced subgraphs. Unlike the lists for the other four basic classes, the one for this class of graphs is finite.

All graphs considered in this paper are finite and have no loops or multiple edges. For a graph G we denote its vertex set by $V(G)$ and its edge set by $E(G)$. The complement of G is denoted by $\overline{G}$. A *clique* in a graph G is a set of vertices all pairwise adjacent and a *stable set* is a clique in $\overline{G}$. For $A \subseteq V(G)$, we denote the subgraph of G induced on A by $G|A$, sometimes further abbreviating $G|\{u, v, w\}$ by $G|uvw$. The notation $G \cong H$ means G is isomorphic to H . For $v \in V(G)$, we denote the set of neighbors of v in G by $N_G(v)$ and for $X \subseteq V(G)$, we denote by $N_X(v)$ the set of neighbors of v in $G|X$.

Let $X, Y \subseteq V(G)$ with $X \cap Y = \emptyset$. We say that X and Y are *complete* to each other if every vertex of X is adjacent to every vertex of Y , and we say that they are *anticomplete* if no vertex of X is adjacent to a member of Y . For an integer $i \geq 0$, let P_i, C_i denote the path and cycle with i edges, respectively.

For integers $a, b \geq 0$, let $M_{a,b}$ be the graph on $2a + b$ vertices consisting of the disjoint union of a edges and b isolated vertices. We say that a graph G is *semi-matched* if it is isomorphic to $M_{a,b}$ for some $a, b \geq 0$ and we say that it is *matched* if in addition $b = 0$. Similarly, we say that G is *semi-antimatched* if it is isomorphic to some $\overline{M_{a,b}}$ and *antimatched* if in addition $b = 0$.

Let $A, B \subseteq V(G)$ such that $A \cap B = \emptyset$, A is semi-matched, and B is semi-antimatched. We say that A and B are *aligned* if the following holds:

- for all adjacent $u, v \in A$ and all $w \in B$, w is adjacent to exactly one of u and v
- for all $u \in A$ and non-adjacent $x, y \in B$, u is adjacent to exactly one of x and y .

A graph G is *split* if its vertex set $V(G)$ can be partitioned into a clique and a

stable set. A graph G is *double-split* if its vertex set $V(G)$ can be partitioned into two sets, A and B , such that the following holds:

- $G|A$ is matched,
- $G|B$ is antimatched, and
- A and B are aligned.

It is easy to see that every split graph is an induced subgraph of many double-split graphs. Also, every induced subgraph of a split graph is also split. Split graphs have a well-known forbidden induced subgraphs characterization:

Theorem 6.1.1. *[Foldes and Hammer [FH77]] A graph is split if and only if it does not contain C_4 , $\overline{C_4}$, or C_5 as an induced subgraph.*

In this paper we consider a class of graphs that includes both split and double-split graphs. We say a graph G is *doubled* if there exists a double-split graph H that contains G as an induced subgraph. Notice that a graph G is double-split if and only if $\overline{G}$ is double-split, and hence a graph G is doubled if and only if $\overline{G}$ is doubled. The main result of this paper is the following:

Theorem 6.1.2. *A graph is doubled if and only if it does not contain any graphs in $\mathcal{F}$, the family of graphs illustrated in Figure 6.1.*

It follows that $\mathcal{F}$ is the list of minimal forbidden induced subgraphs for double-split graphs. The idea for our proof of 6.1.2 is as follows. To prove the "if" part of 6.1.2, we assume that G is not split, hence contains one of C_4 , $\overline{C_4}$, and C_5 . Since C_5 is in $\mathcal{F}$ and the class of doubled graphs is self complementary, we may assume that G has C_4 as an induced subgraph. However, since C_4 is a doubled graph in two different ways (all four vertices can appear on the anti-matched side or 2 vertices can appear on the matched side and the other 2 vertices on the semi-antimatched side),

there is no easy procedure to partition the remaining vertices of the graph. To avoid this obstacle, we introduce another class of graphs that lies in between the class of split graphs and the class of doubled graphs. In section 2, we find the forbidden induced subgraph characterization for this class and we use this characterization to prove 6.1.2 in section 3.

6.2 Almost-split graphs

We say a graph G is *almost-split* if G is doubled and there exists $v \in V(G)$ such that $G|(V(G) \setminus \{v\})$ is split. In other words, G is almost-split if there is at most one pair matched or antimatched. Note that every split graph is almost-split and every almost-split graph is doubled. In this section we present the list of forbidden induced subgraphs for the class of almost-split graphs.

Theorem 6.2.1. *A graph is almost-split if and only if it does not contain any graphs in the circus, the list of graphs illustrated in Figure 6.2 along with their complements.*

Proof. The “only if” part is clear, as it is easy to check that none of the graphs in the circus are almost-split. For the “if” part, suppose that G does not contain any graphs in the circus. By 6.1.1, we may assume that G contains C_4 or $\overline{C_4}$ since split graphs are almost-split. Furthermore, since the statement is self-complementary, we may assume that G contains C_4 . Let $a, b, c, d \in V(G)$ be such that $G|abcd \cong C_4$ and a is adjacent to b and d . Let $S = \{a, b, c, d\}$.

Since $W_4 \cong \overline{M_{2,1}}$ is in the circus, it follows that for all $v \in V(G)$, v is not complete to S . For $0 \leq i \leq 3$, let $A_i \subseteq V(G) \setminus S$ denote the set of vertices that have i neighbors in S . Our goal is to show that there exist adjacent $x, y \in S$ such that:

- $A_0 \cup A_1 \cup A_2 \cup \{x, y\}$ contains only one edge (namely xy), and
- $A_3 \cup (S \setminus \{x, y\})$ is a clique, and

- every vertex of $A_3 \cup (S \setminus \{x, y\})$ is adjacent to exactly one of x and y .

(1) If $A_2 \neq \emptyset$, then there exist $x, y \in S$ such that A_2 is complete to $\{x, y\}$.

Moreover, A_2 is a stable set.

Let $A_{ab} \subseteq A_2$ be those vertices that are adjacent to a and b , and define $A_{ac}, A_{ad}, A_{bc}, A_{bd}, A_{cd}$ similarly. First suppose that $u \in A_{ac} \cup A_{bd}$; then $G|abcd u \cong K_{2,3}$. Hence, both A_{ac} and A_{bd} are empty. Next suppose there exists $u \in A_{ab}$ and $v \in A_{bc}$. Then either $G|abcd uv \cong \text{tent}_2$ or $G|acd uv \cong C_5$, depending on the adjacency between u and v . Therefore, at least one of A_{ab} and A_{bc} is empty, and from symmetry the same is true for the pairs $\{A_{bc}, A_{cd}\}$, $\{A_{cd}, A_{ad}\}$, and $\{A_{ab}, A_{ad}\}$. We claim that at least one of A_{ab} and A_{cd} is empty. For suppose $u \in A_{ab}$ and $v \in A_{cd}$. Then $G|abcd uv \cong \overline{C_6}$ or $G|abcd uv \cong \overline{\text{domino}}$, depending on the adjacency between u and v . Similarly, at least one of A_{bc} and A_{ad} is empty. We conclude that at most one of $A_{ab}, A_{ac}, A_{ad}, A_{bc}, A_{bd}$, and A_{cd} is non-empty. Finally suppose that $u, v \in A_2$ are adjacent. Then $G|abcd uv \cong \overline{\text{watch}}$. Hence, A_2 is a stable set. This proves (1).

(2) There exist adjacent $x, y \in S$ such that $N_S(A_1) \subseteq \{x, y\}$. Moreover, if $A_2 \neq \emptyset$, then $N_S(A_1) \subseteq N_S(A_2)$.

Let $A_a \subseteq A_1$ be those vertices that are adjacent to a , and define A_b, A_c and A_d similarly. We show that at least one of A_a and A_c is empty. For suppose that $u \in A_a$ and $v \in A_c$. Then either $G|abcd uv \cong \text{watch}$ or $G|abc uv \cong C_5$, depending on the adjacency between u and v . Similarly, at least one of A_b and A_d is empty. This proves the first part of (2).

Next, let $u \in A_1$ and $v \in A_2$. Suppose that $N_S(A_1) \not\subseteq N_S(A_2)$. From symmetry, we may assume that $u \in A_a$ and $v \in A_{bc}$. But then either $G|abcd uv \cong \overline{\text{tent}_1}$ or $G|acd uv \cong C_5$, depending on the adjacency between u and v . This proves (2).

(3) $A_0 \cup A_1 \cup A_2$ is a stable set.

First, let $u, v \in A_0$ and suppose that they are adjacent. Then $G|abcd uv \cong \text{TV}$. Hence, A_0 is a stable set. Next, suppose $u, v \in A_1$ and suppose that they are adjacent.

If u, v have a common neighbor in S then $G|abcdv \cong \text{fish}$. If u, v have different neighbors in S , then by (2) their neighbors are adjacent and so $G|abcdv \cong \text{domino}$. This proves that A_1 is a stable set. Recall that A_2 is a stable set by (1).

Now we show that A_0, A_1 , and A_2 are pairwise anticomplete to each other. Let $u \in A_0, v \in A_1$ and suppose that u and v are adjacent. Then $G|abcdv \cong \text{flag}$. Next, let $u \in A_0$ and $v \in A_2$ and again suppose that u and v are adjacent. Then $G|abcdv \cong \overline{\text{tent}_2}$. Finally, let $u \in A_1$ and $v \in A_2$ and suppose that they are adjacent. Then $G|abcdv \cong \text{tent}_1$. Therefore, we have shown that $A_0 \cup A_1 \cup A_2$ is stable. This proves (3).

(4) *There exist adjacent $x, y \in S$ such that A_3 is complete to x, y . Moreover, for all $u \in A_1 \cup A_2$ and $v \in A_3$, $N_S(u) \subseteq N_S(v)$.*

Let $A_{abc} \subseteq A_3$ be the set of vertices that are adjacent to a, b and c , and define A_{abd}, A_{acd} and A_{bcd} similarly. We claim that at least one of A_{abc} and A_{acd} is empty. For suppose that $u \in A_{abc}$ and $v \in A_{acd}$. Then either $G|(S \cup \{u, v\}) \cong \overline{\text{TV}}$ or $G|acd v \cong W_4$, depending on the adjacency between u and v . This proves the claim. By a similar argument, at least one of A_{abd} and A_{bcd} is empty. Therefore, there exist (at least) 2 adjacent vertices of S that are complete to A_3 .

Next, let $u \in A_1 \cup A_2$ and $v \in A_3$ and suppose that $N_S(u) \not\subseteq N_S(v)$. From symmetry, we may assume that $v \in A_{abc}$. If $u \in A_1$, then $u \in A_d$ and so either $G|abcdv \cong \overline{\text{fish}}$ or $G|acd v \cong K_{2,3}$. So we may assume that $u \in A_2$. Again from symmetry, we may assume that $u \in A_{cd}$. But then either $G|abcdv \cong \overline{\text{flag}}$ or $\overline{P_5}$, depending on the adjacency between u and v . This proves (4).

(5) *A_3 is a clique.*

Let $u, v \in A_3$ and suppose that they are not adjacent. By (4), there exist adjacent $x, y \in S$ such that A_3 is complete to $\{x, y\}$, and from symmetry we may assume $\{x, y\} = \{a, b\}$. First suppose that $u, v \in A_{abc}$. Then $G|acd v \cong K_{2,3}$. Therefore, A_{abc} is a clique, and similarly so is A_{abd} . Next suppose that $u \in A_{abc}$ and $v \in A_{abd}$.

Then $G|abcduv \cong \overline{P_5}$. Hence, A_3 is a clique, and this proves (5).

From (1), (2), and (4), it follows that there exist adjacent $x, y \in S$ such that $A_3 \cup A_2$ is complete to $\{x, y\}$ and $N_S(A_1) \subseteq \{x, y\}$. From symmetry, we may assume that $\{x, y\} = \{a, b\}$. Hence, $A_0 \cup A_1 \cup A_2$ is anticomplete to $\{c, d\}$. Therefore, by (3), $A_0 \cup A_1 \cup A_2 \cup \{c, d\}$ contains exactly one edge (namely cd). By (4) and (5), $A_3 \cup \{a, b\}$ is a clique. Also, since every member of A_3 is adjacent to exactly 3 members of S , it follows that for all $u \in A_3 \cup \{a, b\}$, u is adjacent to exactly one of c, d . Hence, we have shown that G is almost-split and this proves 6.2.1. $\square$

6.3 Excluding 6 graphs

In the previous section, we have seen the 12 minimal forbidden induced subgraphs (up to taking complements) for almost-split graphs. Six of them are doubled and the other six are not. In this section, we prove that if a graph contains one of these six doubled graphs but no graphs in $\mathcal{F}$, then it is doubled.

Theorem 6.3.1. *A graph containing $M_{2,1}$ but no graphs in $\mathcal{F}$ is doubled.*

Proof. Let G be a graph containing $M_{2,1}$ but no graphs in $\mathcal{F}$. Let $G|abcde \cong M_{2,1}$, where bc and de are the two edges; let $S = \{a, b, c, d, e\}$. For $0 \leq i \leq 4$, let $A_i \subseteq V(G) \setminus S$ denote the set of vertices that have i neighbors in $\{b, c, d, e\}$. Our goal is to show the following:

- $A_1 = A_3 = A_4 = \emptyset$, and
- $G|(A_0 \cup S)$ is semi-matched, and
- $G|A_2$ is semi-antimatched, and
- $A_0 \cup S$ and A_2 are aligned.

Together, these statements imply that G is doubled.

(1) $A_1 = A_3 = A_4 = \emptyset$. Also, if $v \in A_2$, then v is adjacent to exactly one of b and c , and to exactly one of d and e .

If $v \in A_1$, then $G|abcdev \cong F_7$ or F_8 , depending on the adjacency between v and a . Therefore A_1 is empty. If $v \in A_3$, then $G|abcdev \cong F_9$ or F_{10} , depending on the adjacency between v and a . Therefore A_3 is empty. And if $v \in A_4$, then $G|abcdev \cong F_{11}$ or $G|abcdev \cong F_{12}$, depending on the adjacency between v and a . Therefore A_4 is empty.

Next, let $v \in A_2$. If v is adjacent to b and c , then $G|bcdev \cong \overline{K_{2,3}}$. By symmetry, v is not adjacent to both of d and e . Hence, v is adjacent to exactly one of b and c and to exactly one of d and e . This proves (1).

(2) $G|(A_0 \cup S)$ is semi-matched.

First, we claim that at most one vertex $x \in A_0$ is adjacent to a , and if such a vertex x exists, then x is not adjacent to any other vertices in A_0 . For suppose there are two vertices $x, y \in A_0$, both adjacent to a . If x and y are adjacent, then $G|abcxy \cong \overline{K_{2,3}}$, and if they are not adjacent, $G|abcdxy \cong F_7$. So there is at most one vertex in A_0 adjacent to a . Moreover if there is a vertex $x \in A_0$ adjacent to a , x is not adjacent to any other vertex $y \in A_0$ since otherwise $G|abcdxy \cong F_7$. This proves the claim.

To prove (2), it is enough to show that there do not exist vertices $u, v, w \in A_0 \cup \{a, b, c, d, e\}$ such that $G|uvw \cong C_3$ or $G|uvw \cong P_2$. If at least one of u, v, w is a member of S , then $G|uvw$ cannot be isomorphic to C_3 nor P_2 by the claim. So we may assume $u, v, w \in A_0$. But now if $G|uvw \cong C_3$, then $G|bcuvw \cong \overline{K_{2,3}}$ and if $G|uvw \cong P_2$, then $G|bcduvw \cong F_7$. This proves (2).

(3) Let $u, v \in A_2$ be non-adjacent. Then $N_{\{b,c,d,e\}}(u)$ is disjoint from $N_{\{b,c,d,e\}}(v)$. Moreover, exactly one of u and v is adjacent to a .

From (1) and by symmetry, we may assume that u is adjacent to b and d . Suppose that v is also adjacent to b and d . Then $G|bcdev \cong \text{watch}$. Next, suppose that v

is adjacent to b and e (or c and d). Then $G|bdeuv$ (or $G|bcduv$) is isomorphic to C_5 . Consequently, v is adjacent to c and e .

Moreover, if u and v are both adjacent to a , then $G|abcuv \cong C_5$ and if u and v are both non-adjacent to a , then $G|abcdeuv \cong F_{17}$. Hence, exactly one of u and v is adjacent to a . This proves (3).

(4) $G|A_2$ is semi-antimatched.

It follows easily from (3) that there is no stable set of size 3 in A_2 . Therefore, it is enough to show that there do not exist vertices $u, v, w \in A_2$ such that $G|uvw$ contains exactly one edge (say uv). For contradiction, suppose that such u, v, w exist. From (3) and by symmetry, we may assume that $\{u, v\}$ is complete to $\{b, d\}$, w is complete to $\{c, e\}$, and $N_{\{u, v, w\}}(a)$ is either $\{u, v\}$ or $\{w\}$. In the first case, $G|acuvw \cong \overline{K_{2,3}}$ and in the second case, $G|abuvw \cong \overline{K_{2,3}}$. Therefore, there do not exist $u, v, w \in A_2$ such that $G|uvw$ contains exactly one edge, and this proves (4).

It remains to show that $G|(A_0 \cup S)$ and $G|A_2$ are aligned. In (3), we have shown that for all non-adjacent $u, v \in A_2$ and all $w \in A_0 \cup S$, w is adjacent to exactly one of u and v . Hence, it suffices to show that for all $u \in A_2$ and all adjacent $v, w \in A_0 \cup S$, u is adjacent to exactly one of v, w . So suppose that for some u, v, w as above, u is adjacent to both of v, w . Let $x, y \in A_0 \cup S$ be adjacent such that $\{x, y\}$ is disjoint from $\{v, w\}$ (such x, y exist since $A_0 \cup S$ contain at least two edges). Then $G|uvwxy \cong \overline{K_{2,3}}$. Next, suppose that for some u, v, w as above, u is non-adjacent to both of v, w . Note that by (1), $\{v, w\}$ is disjoint from $\{b, c, d, e\}$. By (1) and without loss of generality, we may assume that u is adjacent to b and d . But then $G|bceuvw \cong F_7$. Therefore G is doubled and this proves 6.3.1. $\square$

Theorem 6.3.2. *A graph containing P_5 but no graphs in $\mathcal{F}$ is doubled.*

Proof. Let G be a graph containing P_5 but no graphs in $\mathcal{F}$. Let $G|abcdef \cong P_5$ where ab, bc, cd, de , and ef are the five edges. By 6.3.1, we may assume that G or $\overline{G}$ does not contain $M_{2,1}$. Let $S = \{a, b, c, d, e, f\}$. For $0 \leq i \leq 4$, let $A_i \subseteq V(G) \setminus S$

denote the set of vertices that have i neighbors in $\{b, c, d, e\}$. Our goal is to show the following:

- $A_0 = A_2 = A_4 = \emptyset$, and
- $G|(A_1 \cup \{a, c, d, f\})$ is semi-matched, and
- $G|(A_3 \cup \{b, e\})$ is semi-antimatched, and
- $G|(A_1 \cup \{a, c, d, f\})$ and $G|(A_3 \cup \{b, e\})$ are aligned.

Together, these statements imply that G is doubled.

(1) $A_0 = A_2 = A_4 = \emptyset$.

First suppose $v \in A_0$. If v is non-adjacent to a , then $G|abdev \cong M_{2,1}$. So we may assume that v is adjacent to a and similarly v is adjacent to f ; but then $G|abcdefv \cong F_{13}$. Therefore A_0 is empty.

Next, suppose $v \in A_4$. If v is not adjacent to both of a and f , then $G|abcdefv \cong F_{14}$. If v is adjacent to exactly one of a and f , then $G|abcdefv \cong F_{15}$. So we may assume that v is adjacent to both a and f ; but then $G|abcdefv \cong F_{16}$. Therefore A_4 is empty.

Finally suppose $v \in A_2$. Let $A_{bc} \subseteq A_2$ be those vertices that are adjacent to b and c , and define A_{bd} , A_{be} , A_{cd} , A_{ce} , A_{de} similarly. Suppose $v \in A_{bc}$. If v is adjacent to f , then $G|cdefv \cong C_5$; otherwise, $G|bcefv \cong \overline{K_{2,3}}$. Therefore A_{bc} is empty and by symmetry, A_{de} is empty. Next, suppose $v \in A_{bd}$. If v is not adjacent to a , then $G|abcdev \cong \text{watch}$ and if v is not adjacent to f , then $G|bcdefv \cong \text{flag}$. So we may assume that v is adjacent to both of a and f ; but then $G|abdefv \cong \text{fish}$. Therefore A_{bd} is empty and by symmetry, A_{ce} is empty. Next, suppose $v \in A_{be}$. Then $G|bcdev \cong C_5$. Therefore A_{be} is empty. So we may assume that $v \in A_{cd}$. If v is adjacent to both of a and f , then $G|abcefv \cong \text{flag}$. If v is adjacent to a and not to f , then $G|abcefv \cong \text{TV}$, and if v is adjacent to f and not to a , then $G|abdefv \cong \text{TV}$. So we may assume that

v is not adjacent to either a or f ; but then $G|abefv \cong M_{2,1}$. Therefore $A_2 = \emptyset$ and this proves (1).

(2) If $v \in A_1$, then v is adjacent to either b or e and is not adjacent to both a and f . Moreover, A_1 is a stable set.

Let $A_b \subseteq A_1$ be those vertices that are adjacent to b , and define A_c, A_d, A_e similarly. Suppose $v \in A_c$. If v is adjacent to a , then $G|abcdev \cong \text{flag}$; otherwise, $G|abdev \cong M_{2,1}$. Therefore A_c is empty, and by symmetry, A_d is empty. Suppose $v \in A_b$. If v is adjacent to a , then $G|abdev \cong \overline{K_{2,3}}$ and if v is adjacent to f , then $G|acdfv \cong M_{2,1}$. Therefore v is anticomplete to $\{a, f\}$, and similarly, every vertex in A_e is anticomplete to $\{a, f\}$.

Next, suppose that $u, v \in A_b$ are adjacent. Then $G|bdeuv \cong \overline{K_{2,3}}$. Therefore A_b is a stable set and similarly, so is A_e . Finally, suppose that $u \in A_b$ and $v \in A_e$ are adjacent. Then $G|acduv \cong M_{2,1}$. Therefore $A_b \cup A_e = A_1$ is a stable set, and this proves (2).

(3) If $v \in A_3$, then $N_{\{b,c,d,e\}}(v)$ is either $\{b, c, e\}$ or $\{b, d, e\}$. Moreover, if $u, v \in A_3$ are not adjacent, then $N_{\{u,v\}}(c) \neq N_{\{u,v\}}(d)$ and $|N_{\{u,v\}}(a)| = |N_{\{u,v\}}(f)| = 1$.

Let $A_{bcd} \subseteq A_3$ be those vertices that are adjacent to b, c , and d , and define $A_{bce}, A_{bde}, A_{cde}$ similarly. Suppose $v \in A_{cde}$. If v is not adjacent to a , then $G|abdev \cong \overline{K_{2,3}}$, and if v is adjacent to a and f , then $G|abcefv \cong \text{fish}$. So we may assume that v is adjacent to a but not to f ; but then $G|abcefv \cong \text{flag}$. Therefore A_{cde} is empty and similarly, so is A_{bcd} .

Now suppose $u, v \in A_{bce}$ are not adjacent; then $G|cdeuv \cong K_{2,3}$. Therefore A_{bce} is a clique and similarly, so is A_{bde} . Suppose $u \in A_{bce}$ and $v \in A_{bde}$ are not adjacent. If u, v are both adjacent to a , then $G|abdeuv \cong \overline{\text{flag}}$, and if u, v are both non-adjacent to a , then $G|abcdeuv \cong \overline{F_{18}}$. Therefore exactly one of u and v is adjacent to a and similarly, exactly one of u and v is adjacent to f . This proves (3).

(4) $G|(A_3 \cup \{b, e\})$ is semi-antimatched.

It is enough to show that no set of three vertices $\{u, v, w\} \subseteq A_3 \cup \{b, e\}$ contains fewer than two edges. By (3), it is obvious that there are no stable sets of size 3 in $G|(A_3 \cup \{b, e\})$. Suppose $\{u, v, w\}$ contains exactly one edge uv . From (3) and symmetry, we may assume $u, v \in A_{bce}$ and $w \in A_{bde}$. But then $G|cdeuvw \cong \overline{\text{flag}}$. This proves (4).

From (2) and (4), we have a candidate of a partition for G to be doubled. The subgraph $G|(A_1 \cup \{a, c, d, f\})$ contains only one edge (namely cd) and $G|(A_3 \cup \{b, e\})$ is semi-antimatched. Every $v \in A_3 \cup \{b, e\}$ has exactly one neighbor in $\{c, d\}$ and from (3), for every non adjacent pair $u, v \in A_3 \cup \{b, e\}$, $N_{\{u, v\}}(c) \neq N_{\{u, v\}}(d)$. Also by (3), if $u, v \in A_3 \cup \{b, e\}$ are nonadjacent, $|N_{\{u, v\}}(a)| = |N_{\{u, v\}}(f)| = 1$. Moreover, for $w \in A_1$, either $G|abcdeuw \cong P_5$ or $G|bcdefw \cong P_5$, and so $|N_{\{u, v\}}(w)| = 1$ for every $w \in A_1$, by an analogous argument to the one above. Therefore, $G|(A_3 \cup \{b, e\})$ and $G|(A_1 \cup \{a, c, d, f\})$ are aligned and so G is doubled; this proves 6.3.2. $\square$

Theorem 6.3.3. *A graph containing $\overline{C_6}$ but no graphs in $\mathcal{F}$ is doubled.*

Proof. Let G be a graph containing $\overline{C_6}$ but no graphs in $\mathcal{F}$. Let $G|abcdef \cong \overline{C_6}$ where $\{a, c, e\}$ and $\{b, d, f\}$ are the two triangles and the remaining edges are ad, be , and cf . Let $S = \{a, b, c, d, e, f\}$. By 6.3.1, we may assume G or $\overline{G}$ does not contain $M_{2,1}$. For $0 \leq i \leq 6$, let $A_i \subseteq V(G) \setminus S$ denote the set of vertices that have i neighbors in S . Our goal is to show that $A_i = \emptyset$ unless $i = 2$ and 4 vertices of S induce antimatching side and the rest of vertices (two in S together with vertices in A_2) induce matching side so that G is doubled.

(1) $A_i = \emptyset$ for $i = 0, 1, 3, 4, 5, 6$.

If $v \in A_0$, then $G|(S \cup \{v\}) \cong \overline{F_{19}}$, so A_0 is empty. Also, if $v \in A_6$, then $G|(S \cup \{v\}) \cong \overline{F_{17}}$, and so A_6 is empty.

Next, suppose $v \in A_1$. From symmetry, we may assume $N_S(v) = \{a\}$. Then $G|(S \cup \{v\}) \cong F_{18}$ and therefore A_1 is empty.

Next, suppose $v \in A_3$. From symmetry, we may assume $N_S(v)$ is one of $\{a, b, c\}, \{a, b, d\}, \{a, c, e\}$.

If $N_S(v) = \{a, b, c\}$, then $G|abdev \cong K_{2,3}$ and if $N_S(v) = \{a, b, d\}$, then $G|abcfv \cong C_5$. So we may assume that $N_S(v) = \{a, c, e\}$; but then $G|abcdev \cong \overline{\text{watch}}$ and so A_3 is empty.

Next, suppose $v \in A_4$. From symmetry, we may assume $N_S(v)$ is one of $\{a, b, c, d\}$, $\{a, b, c, e\}$, and $\{a, b, d, e\}$. If $N_S(v) = \{a, b, c, d\}$, then $G|bcefv \cong K_{2,3}$ and if $N_S(v) = \{a, b, c, e\}$, then $G|acdefv \cong \overline{\text{watch}}$. So we may assume that $N_S(v) = \{a, b, d, e\}$; but then $G|abcfv \cong C_5$, and so A_4 is empty.

Finally, suppose $v \in A_5$. From symmetry, we may assume $N_S(v) = \{b, c, d, e, f\}$. Then $G|bcefv \cong \overline{M_{2,1}}$. Therefore A_5 is empty and this proves (1).

For $u, v \in S$, let $A_{uv} \subseteq A_2$ be those vertices that are adjacent to u and v .

$$(2) \ A_{ab} = A_{bc} = A_{cd} = A_{de} = A_{ef} = A_{fa} = \emptyset.$$

Suppose $v \in A_{ab}$. Then $G|abcfv \cong C_5$. Therefore A_{ab} is empty and similarly, so are A_{bc} , A_{cd} , A_{de} , A_{ef} , and A_{fa} .

$$(3) \ \text{For every } x, y \in S, \ A_{xy} \text{ is a stable set.}$$

Suppose $u, v \in A_{ac}$ are adjacent. Then $G|acdfuv \cong \overline{\text{watch}}$. Therefore A_{ac} is a stable set and similarly so are A_{ae} , A_{ce} , A_{bd} , A_{bf} , and A_{df} . Suppose $u, v \in A_{ad}$ are adjacent. Then $G|acdfuv \cong \overline{\text{watch}}$. Therefore A_{ad} is a stable set and similarly so are A_{be} and A_{cf} . This proves (3).

$$(4) \ \text{If } A_{ac} \neq \emptyset, \text{ then } A_{ae} = A_{ce} = A_{bd} = A_{bf} = \emptyset.$$

Suppose $u \in A_{ac}$, and $v \in A_{ae}$. If u and v are adjacent, then $G|aceuv \cong \overline{M_{2,1}}$, and otherwise $G|bcduv \cong M_{2,1}$. Therefore if A_{ac} is not empty, then $A_{ae} = \emptyset$ and similarly, $A_{ce} = \emptyset$.

Now suppose $v \in A_{bd}$. If u and v are adjacent, then $G|cdfuv \cong C_5$, and otherwise $G|bcefv \cong \text{watch}$. Therefore if A_{ac} is not empty, then $A_{bd} = \emptyset$ and similarly, $A_{bf} = \emptyset$. This proves (4).

$$(5) \ \text{If } A_{ad} \neq \emptyset, \text{ then } A_{ce} = A_{bf} = \emptyset.$$

Suppose $u \in A_{ad}$ and $v \in A_{ce}$. If u and v are adjacent, then $G|cdfuv \cong C_5$, and

otherwise $G|cdeuv \cong \overline{K_{2,3}}$. Therefore if A_{ad} is not empty, then A_{ce} is empty and similarly, A_{bf} is empty as well. This proves (5).

(6) If $u \in A_{ad}$, then $N_G(u) \setminus S \subseteq A_{be} \cup A_{cf}$.

Suppose $u \in A_{ad}$. Then from (5), $A_{ce} = A_{bf} = \emptyset$, and from (3), u has no neighbors in A_{ad} . Now suppose $v \in A_{ac}$ is adjacent to u . Then $G|cdfuv \cong C_5$. Therefore u is anticomplete to A_{ac} , and similarly, u is anticomplete to A_{ae} , A_{bd} , and A_{df} as well. Therefore $N_G(u) \setminus S \subseteq A_{be} \cup A_{cf}$, and this proves (6).

(7) If there are adjacent vertices $u \in A_{ad}$ and $v \in A_{be} \cup A_{cf}$, then $V(G) = S \cup \{u, v\}$ and G is doubled.

From symmetry, we may assume $v \in A_{cf}$ is adjacent to $u \in A_{ad}$. We know that $A_{ae} \cup A_{ce} \cup A_{bd} \cup A_{bf} = \emptyset$ by (5). Suppose $w \in A_{ac} \cup A_{df}$. Then from (6), $\{u, v\}$ is anticomplete to w and so $G|beuvw \cong M_{2,1}$. Therefore $A_{ac} \cup A_{ae} \cup A_{ce} \cup A_{bd} \cup A_{bf} \cup A_{df} = \emptyset$.

Next, suppose $w(\neq v) \in A_{cf}$. From (3), w is not adjacent to v . If w is adjacent to u , then $G|abfuuv \cong \text{watch}$, and otherwise $G|beuvw \cong M_{2,1}$. Therefore $A_{cf} = \{v\}$ and similarly, $A_{ad} = \{u\}$.

Now suppose $w \in A_{be}$. If w is anticomplete to $\{u, v\}$, then $G|beuvw \cong \overline{K_{2,3}}$. Therefore w is adjacent to at least one of $\{u, v\}$ and by the same logic as above, $A_{be} = \{w\}$. If w is adjacent to exactly one of u and v (say u), then $G|abcuvw \cong \text{flag}$. So we may assume that w is adjacent to both u and v ; but then $G|(S \cup \{u, v, w\}) \cong F_{23}$. Therefore $A_{be} = \emptyset$. But then $V(G) = S \cup \{u, v\}$. Since $G|uvbe$ is matched, $G|abcf$ is antimatched, and the two subgraphs are aligned, it follows that G is doubled. This proves (7).

(8) If $v \in A_{ac}$, then $N_G(v) \setminus S \subseteq A_{df}$.

Suppose $v \in A_{ac}$. From (3), v has no neighbors in A_{ac} . From (4), $A_{ae} = A_{ce} = A_{bd} = A_{bf} = \emptyset$ and from (5), $A_{be} = \emptyset$. Finally, from (6), v is anticomplete to $A_{ad} \cup A_{cf}$. Therefore $N_G(v) \setminus S \subseteq A_{df}$, and this proves (8).

(9) If there are adjacent vertices $u \in A_{ac}$ and $v \in A_{df}$, then $V(G) = S \cup \{u, v\}$ and G is doubled.

From (4) and (5), $A_{ae} = A_{ce} = A_{bd} = A_{bf} = A_{be} = \emptyset$. If $w \in A_{ad} \cup A_{cf}$, then from (8), w is anticomplete to $\{u, v\}$ and so $G|beuvw \cong M_{2,1}$. Therefore $A_{ad} \cup A_{cf}$ is empty and $V(G) \setminus S = A_{ac} \cup A_{df}$.

Now suppose $w(\neq v) \in A_{df}$. From (3), w is not adjacent to v . If w is adjacent to u , then $G|cdeuvw \cong \text{flag}$, and otherwise $G|beuvw \cong M_{2,1}$. Therefore $A_{df} = \{v\}$ and similarly, $A_{ac} = \{u\}$. Hence, $V(G) = S \cup \{u, v\}$. Since $G|uvbe$ is matched, $G|abcf$ is antimatched, and the two subgraphs are aligned, it follows that G is doubled. This proves (9).

(10) If $G|(V(G) \setminus S)$ is a stable set, then G is doubled.

Suppose $G|(V(G) \setminus S)$ is a stable set. First, suppose $A_{ac} \cup A_{ce} \cup A_{ae} \cup A_{bd} \cup A_{bf} \cup A_{df} \neq \emptyset$. From symmetry, we may assume $A_{ac} \neq \emptyset$. Then from (4), $A_{ae} = A_{ce} = A_{bd} = A_{bf} = \emptyset$ and from (5), $A_{be} = \emptyset$. Therefore every vertex in $V(G) \setminus S$ has exactly one neighbor in $\{a, f\}$ and exactly one neighbor in $\{c, d\}$. Now it is easy to see that G is doubled with $G|acdf$ as the antimatched part.

Therefore we may assume $A_{ac} \cup A_{ce} \cup A_{ae} \cup A_{bd} \cup A_{bf} \cup A_{df} = \emptyset$. Suppose all three of the sets A_{ad} , A_{be} , and A_{cf} are not empty. Then for $u \in A_{ad}$, $v \in A_{be}$, and $w \in A_{cf}$, $G|afuvw \cong M_{2,1}$. Therefore from symmetry, we may assume A_{be} is empty. Now again, every vertex in $V(G) \setminus S$ has exactly one neighbor in $\{a, f\}$ and exactly one neighbor in $\{c, d\}$, so G is doubled with $G|acdf$ as the antimatched part. This proves (10).

By (10), we may assume that $G|V(G) \setminus S$ contains an edge uv . From symmetry, we may assume $u \in A_{ad}$ or $u \in A_{ac}$. If $u \in A_{ad}$, then by (6) and (7), $v \in A_{be} \cup A_{cf}$ and G is doubled. So we may assume that $u \in A_{ac}$; but then by (8) and (9), $v \in A_{df}$ and G is doubled. This proves 6.3.3. $\square$

Theorem 6.3.4. *A graph containing $\overline{\text{domino}}$ but no graphs in $\mathcal{F}$ is doubled.*

Proof. Let G be a graph containing $\overline{\text{domino}}$ but no graphs in $\mathcal{F}$. By 6.3.1, 6.3.2, and 6.3.3, we may assume that G does not contain $M_{2,1}$, P_5 , C_6 , or their complements as induced subgraphs. Let $G|abcdef \cong \overline{\text{domino}}$, where $ab, bc, ca, bd, ce, de, df$, and ef are the edges; let $S = \{a, b, c, d, e, f\}$. For $0 \leq i \leq 4$, let $A_i \subseteq V(G) \setminus S$ denote the set of vertices that have i neighbors in $\{b, c, d, e\}$. Our goal is to show the following:

- $A_0 = A_1 = A_3 = A_4 = \emptyset$, and
- $G|(A_2 \cup \{a, f\})$ is a stable set, and
- $G|bcde$ is antimatched, and
- $A_2 \cup \{a, f\}$ and $\{b, c, d, e\}$ are aligned.

Together, these statements imply that G is doubled.

(1) $A_0 = A_1 = A_3 = A_4 = \emptyset$.

Suppose $v \in A_0$. If v is complete to $\{a, f\}$, then $G|abdfv \cong C_5$, and if v is anticomplete to $\{a, f\}$, then $G|acdfv \cong M_{2,1}$. So we may assume that v is adjacent to exactly one of a and f , say a ; but then $G|adev \cong \overline{K_{2,3}}$. Therefore $A_0 = \emptyset$.

Next, suppose $v \in A_1$. From symmetry, we may assume $N_{\{b,c,d,e\}}(v) = \{b\}$. If v is complete to $\{a, f\}$, then $G|acefv \cong C_5$, and if v is anticomplete to $\{a, f\}$, then $G|acdfv \cong M_{2,1}$. Furthermore, if v is adjacent to a but not to f , then $G|abefv \cong \overline{K_{2,3}}$. So we may assume that v is adjacent to f but not to a ; but then $G|bcefv \cong C_5$. Therefore $A_1 = \emptyset$.

Next, suppose $v \in A_3$. From symmetry, we may assume $N_{\{b,c,d,e\}}(v) = \{b, c, e\}$. If v is not adjacent to f , then $G|bcdev \cong \overline{\text{flag}}$, and if v is complete to $\{a, f\}$, then $G|abcfv \cong \overline{\text{watch}}$. So we may assume that v is adjacent to f but not to a ; but then $G|abdefv \cong \overline{\text{fish}}$. Therefore $A_3 = \emptyset$.

Finally, suppose $v \in A_4$. Then $G|bcdev \cong \overline{M_{2,1}}$. Therefore $A_4 = \emptyset$. This proves (1).

For $u, v \in \{b, c, d, e\}$, let $A_{uv} \subseteq A_2$ be those vertices that are adjacent to u and v .

(2) $A_{be} = A_{cd} = \emptyset$. Moreover, $A_2 \cup \{a, f\}$ is a stable set.

Suppose $v \in A_{be} \cup A_{cd}$; then $G|bcdev \cong K_{2,3}$. Therefore $A_{be} = A_{cd} = \emptyset$. Next, suppose $v \in A_{bc}$. If v is adjacent to a , then $G|abcdev \cong \overline{\text{watch}}$ and if v is adjacent to f , then $G|bcdefv \cong \overline{C_6}$. Therefore A_{bc} is anticomplete to $\{a, f\}$, and from symmetry, so is A_{de} .

Now suppose $v \in A_{bd}$. If v is adjacent to a , then $G|acdev \cong C_5$ and if v is adjacent to f , then $G|cdefv \cong C_5$. Therefore A_{bd} is anticomplete to $\{a, f\}$, and from symmetry, so is A_{ce} . It follows that A_2 is anticomplete to $\{a, f\}$. Note that for $v \in A_{bc} \cup A_{de}$, either $G|abcdev \cong \text{domino}$ or $G|bcdefv \cong \text{domino}$, and so by an argument analogous to the one above, we conclude that $A_{bc} \cup A_{de}$ is anticomplete to $A_{bd} \cup A_{ce}$ and that $A_{bc} \cup A_{de}$ is a stable set; hence $A_{bc} \cup A_{de} \cup \{a, f\}$ is a stable set.

It remains to show that $A_{bd} \cup A_{ce}$ is a stable set. For suppose $u, v \in A_{bd}$ are adjacent; then $G|bcdeuv \cong \overline{\text{watch}}$. Therefore A_{bd} is a stable set and from symmetry, so is A_{ce} . Next, suppose $u \in A_{bd}$ and $v \in A_{ce}$ are adjacent; then $G|bcdeuv \cong \overline{C_6}$. Therefore $A_2 \cup \{a, f\}$ is a stable set and this proves (2).

Now $\{b, c, d, e\}$ is anti-matched by definition and $A_2 \cup \{a, f\}$ is a stable set by (2). It remains to show that $A_2 \cup \{a, f\}$ and $\{b, c, d, e\}$ are aligned. Since $A_2 \cup \{a, f\}$ is a stable set, it suffices to show that for all $v \in A_2 \cup \{a, f\}$, v is adjacent to exactly one of b, e and exactly one of c, d . For $v \in \{a, f\}$ this is true by definition, and for $v \in A_2$ this follows from (2). Therefore G is doubled and this proves 6.3.4. $\square$

Theorem 6.3.5. *A graph containing tent_1 but no graphs in $\mathcal{F}$ is doubled.*

Proof. Let G be a graph containing tent_1 but no graphs in $\mathcal{F}$. By 6.3.1, 6.3.2, 6.3.3, and 6.3.4, we may assume that G does not contain $M_{2,1}$, P_5 , C_6 , domino or their complements as induced subgraphs. Let $G|abcdef \cong \text{tent}_1$, where $ab, bc, cd, de, fa, fb, fc$, and fe are the edges; let $S = \{a, b, c, d, e, f\}$. For $0 \leq i \leq 4$, let $A_i \subseteq V(G) \setminus S$ denote the set of vertices that have i neighbors in $\{b, c, d, e\}$. Our goal is to show the

following:

- $A_0 = A_2 = A_4 = \emptyset$, and
- $G|(A_1 \cup \{a, c, d\})$ is semi-matched, and
- $G|(A_3 \cup \{b, e, f\})$ is semi-antimatched, and
- $A_1 \cup \{a, c, d\}$ and $A_3 \cup \{b, e, f\}$ are aligned.

Together, these statements imply that G is doubled.

(1) $A_0 = A_2 = A_4 = \emptyset$.

Suppose $v \in A_0$. If v is adjacent to a , then $G|abcdev \cong P_5$, and if v is not adjacent to a , then $G|abdev \cong M_{2,1}$. Therefore $A_0 = \emptyset$.

Next, suppose $v \in A_4$. If v is adjacent to f , then $G|cdefv \cong \overline{M_{2,1}}$, and if v is not adjacent to f , then $G|bcdefv \cong \overline{P_5}$. Therefore $A_4 = \emptyset$.

Next, we show that $A_2 = \emptyset$. For $u, v \in \{b, c, d, e\}$, let $A_{uv} \subseteq A_2$ be those vertices that are adjacent to u and v . If $v \in A_{be}$, then $G|bcdev \cong C_5$, and so $A_{be} = \emptyset$. Now suppose $v \in A_{bc}$. If v is adjacent to a , then $G|abdev \cong \overline{K_{2,3}}$ and if v is adjacent to f , then $G|bcdefv \cong \overline{\text{watch}}$. So we may assume that v is not adjacent to either a or f ; but then $G|S \cup \{v\} \cong F_{20}$. Therefore $A_{bc} = \emptyset$.

Next, suppose $v \in A_{bd}$. If v is not adjacent to f , then $G|bdefv \cong C_5$, and if v is not adjacent to a , then $G|abcdev \cong \text{watch}$. Hence, we may assume that v is adjacent to both a and f ; but then $G|abdefv \cong \overline{\text{watch}}$. Therefore $A_{bd} = \emptyset$.

Next, suppose $v \in A_{cd}$. If v is adjacent to both a and f , then $G|abcfv \cong \overline{M_{2,1}}$. Next, if v is adjacent to a but not to f , then $G|abcdfv \cong \overline{\text{flag}}$, and if v is adjacent to f but not to a , then $G|abdefv \cong \text{fish}$. So we may assume that v is not adjacent to a or f ; but then $G|abdfv \cong \overline{K_{2,3}}$. Therefore $A_{cd} = \emptyset$.

Next, suppose $v \in A_{ce}$. Then $G|abcdev \cong \text{domino}$ or flag depending on the adjacency between v and a . Therefore $A_{ce} = \emptyset$.

So we may assume that $v \in A_{de}$. If v is adjacent to a , then $G|abcdv \cong C_5$, and if v is not adjacent to a , then $G|abdev \cong \overline{K_{2,3}}$. Therefore $A_2 = \emptyset$ and this proves (1).

(2) A_1 is complete to b .

For $u \in \{b, c, d, e\}$, let $A_u \subseteq A_1$ be those vertices that are adjacent to u . We will show that $A_c = A_d = A_e = \emptyset$.

Suppose $v \in A_c$. If v is adjacent to a , then $G|abcdev \cong \text{flag}$, and if v is not adjacent to a , then $G|abdev \cong M_{2,1}$. Therefore $A_c = \emptyset$.

Next, suppose $v \in A_e$. Then $G|abcdev \cong P_5$ or C_6 depending on the adjacency between v and a . Therefore $A_e = \emptyset$.

Next, suppose $v \in A_d$. If v is adjacent to a , then $G|abcdv \cong C_5$, and if v is adjacent to f but not to a , then $G|abdefv \cong \text{fish}$. So we may assume that v is not adjacent to either a or f ; but then $G|acdefv \cong \text{watch}$. Therefore $A_d = \emptyset$. This completes that proof of (2).

(3) $A_1 \cup \{a\}$ is a stable set.

Suppose $v \in A_1$ and a are adjacent; then $G|abdev \cong \overline{K_{2,3}}$. Therefore A_1 is anticomplete to a . Next, suppose $u, v \in A_1$ are adjacent. Then $G|bdeuv \cong \overline{K_{2,3}}$. Therefore $A_1 \cup \{a\}$ is a stable set and this proves (3).

(4) If $v \in A_3$, then $v \in A_{bce} \cup A_{bde}$.

For $u, v, w \in \{b, c, d, e\}$, let $A_{uvw} \subseteq A_3$ be those vertices that are adjacent to u, v and w .

Suppose $v \in A_{bcd}$. If v is not adjacent to f , then $G|bdefv \cong C_5$, and if v is adjacent to f but not to a , then $G|S \cup \{v\} \cong \overline{F_{21}}$. So we may assume that v is adjacent to both a and f ; but then $G|abdefv \cong \overline{\text{watch}}$. Therefore $A_{bcd} = \emptyset$.

Next, suppose $v \in A_{cde}$. If v is adjacent to f , then $G|cdefv \cong \overline{M_{2,1}}$, and if v is not adjacent to f , then $G|bcdefv \cong \overline{\text{flag}}$. Therefore $A_{cde} = \emptyset$. This proves (4).

(5) $A_3 \cup \{f\}$ is a clique.

Suppose $v \in A_{bde}$ and v is not adjacent to f . Then $G|bcdefv \cong \overline{C_6}$. Next, suppose

$v \in A_{bce}$ not adjacent to f . Then $G|cdefv \cong K_{2,3}$. Therefore A_3 is complete to f .

Next, suppose $u, v \in A_{bde} \cup A_{bce}$ are not adjacent. If $u, v \in A_{bde}$, then $G|bcdv \cong K_{2,3}$ and if $u, v \in A_{bce}$ then $G|cdeuv \cong K_{2,3}$. So we may assume that $u \in A_{bde}$ and $v \in A_{bce}$; but then $G|bcdeuv \cong \overline{C_6}$. Therefore $A_3 \cup \{f\}$ is a clique and this proves (5).

From (2) and (3), it follows that $A_1 \cup \{a, c, d\}$ is semi-matched with one edge (namely, cd). From (4) and (5), $A_3 \cup \{b, e, f\}$ is semi-antimatched with one nonedge (namely, be). Furthermore, it follows by definition and from (2) that for all $u \in A_1 \cup \{a, c, d\}$, u is adjacent to exactly one of b and e . It also follows by definition and from (4) that for all $v \in A_3 \cup \{b, e, f\}$, v is adjacent to exactly one of c and d . Therefore $A_1 \cup \{a, c, d\}$ and $A_3 \cup \{b, e, f\}$ are aligned and this proves 6.3.5. $\square$

Theorem 6.3.6. *A graph containing tent_2 but no graphs in $\mathcal{F}$ is doubled.*

Proof. Let G be a graph containing tent_2 but no graphs in $\mathcal{F}$. By 6.3.1, 6.3.2, 6.3.3, 6.3.4, and 6.3.5, we may assume that G does not contain $M_{2,1}$, P_5 , C_6 , domino, tent_1 or their complements as induced subgraphs. Let $G|abcdef \cong \text{tent}_2$, where $ab, bc, cd, de, fa, fb, fd$, and fe are the edges; let $S = \{a, b, c, d, e, f\}$. First, we show that if $v \in V(G) \setminus S$, then $N_S(v)$ is equal to $\{b, f\}$, $\{d, f\}$, or $\{a, b, d, e, f\}$.

Let A_{bf} be those vertices whose neighbor set in S is $\{b, f\}$ and define A_{df} and A_{abdef} similarly. We also prove that at least one of A_{bf} , A_{df} and A_{abdef} is empty. Then our goal is to show the following:

If $A_{abdef} = \emptyset$, then

- $G|(A_2 \cup \{a, e\})$ is semi-matched, and
- $G|\{b, c, d, f\}$ is antimatched, and
- $A_2 \cup \{a, e\}$ and $\{b, c, d, f\}$ are aligned.

If $A_{df} = \emptyset$, then

- $G|(A_2 \cup \{a, c, d\})$ is semi-matched, and

- $G|(A_5 \cup \{b, e, f\})$ is semi-antimatched, and
- $A_2 \cup \{a, c, d\}$ and $A_5 \cup \{b, e, f\}$ are aligned.

If $A_{bf} = \emptyset$, then

- $G|(A_2 \cup \{b, c, e\})$ is semi-matched, and
- $G|(A_5 \cup \{a, d, f\})$ is semi-antimatched, and
- $A_2 \cup \{b, c, e\}$ and $A_5 \cup \{a, d, f\}$ are aligned.

Together, these statements imply that G is doubled.

(1) For $v \in V(G) \setminus S$, $N_S(v)$ is equal to $\{b, f\}$, $\{d, f\}$, or $\{a, b, d, e, f\}$.

We show that $N_{\{b, c, d\}}(v)$ is equal to $\{b\}$, $\{d\}$, or $\{b, d\}$ and for each case, $N_S(v)$ is equal to $\{b, f\}$, $\{d, f\}$, or $\{a, b, d, e, f\}$, respectively.

First, suppose $N_{\{b, c, d\}}(v) = \emptyset$. If v is complete to $\{a, e\}$, then $G|abcdev \cong C_6$, and if v is adjacent to exactly one of a and e , then $G|abcdev \cong P_5$. So we may assume that v is anticomplete to $\{a, e\}$; but then $G|abdev \cong M_{2,1}$. Therefore $N_{\{b, c, d\}}(v)$ cannot be empty.

Next, suppose $N_{\{b, c, d\}}(v) = \{b\}$. If v is adjacent to e , then $G|bcdev \cong C_5$, and if v is adjacent to a , then $G|abdev \cong \overline{K_{2,3}}$. If v is not adjacent to f , then $G|bcdefv \cong \overline{\text{tent}_1}$. Therefore $N_S(v) = \{b, f\}$. Similarly if $N_{\{b, c, d\}}(v) = \{d\}$, then $N_S(v) = \{d, f\}$.

Next, suppose $N_{\{b, c, d\}}(v) = \{c\}$. If v is complete to $\{a, e\}$, then $G|abcdev \cong \text{domino}$, and if v is adjacent to exactly one of a and e , then $G|abcdev \cong \text{flag}$. So we may assume that v is anticomplete to $\{a, e\}$; but then $G|abdev \cong M_{2,1}$. Therefore $N_{\{b, c, d\}}(v)$ cannot be equal to $\{c\}$.

Next, suppose $N_{\{b, c, d\}}(v) = \{b, c\}$. If v is complete to $\{e, f\}$, then $G|bcdefv \cong \overline{P_5}$. If v is adjacent to e but not to f , then $G|bcdefv \cong \overline{C_6}$, and if v is adjacent to f but not to e , then $G|bcdefv \cong \overline{\text{flag}}$. So we may assume that v is anticomplete to

$\{e, f\}$; but then $G|bcdefv \cong \overline{\text{domino}}$. Therefore $N_{\{b,c,d\}}(v)$ cannot be $\{b, c\}$ and from symmetry, $N_{\{b,c,d\}}(v)$ cannot be $\{c, d\}$.

Next, suppose $N_{\{b,c,d\}}(v) = \{b, d\}$. If v is not adjacent to f , then $G|bcdv \cong K_{2,3}$. If v is anticomplete to $\{a, e\}$, then $G|abcdev \cong \text{watch}$. If v is adjacent to one of a and e , then $G|abcdev \cong \overline{\text{tent}_1}$. Therefore $N_S(v) = \{a, b, d, e, f\}$.

Finally, suppose $N_{\{b,c,d\}}(v) = \{b, c, d\}$. If v is adjacent to f , then $G|bcdv \cong \overline{M_{2,1}}$. If v is not adjacent to a , then $G|abcdv \cong \overline{\text{flag}}$, while if v is adjacent to a , then $G|abcdv \cong \overline{P_5}$. Therefore v cannot be complete to $\{b, c, d\}$.

Together, these statements prove (1).

(2) $A_{bf} \cup A_{df}$ is a stable set, and A_{abdef} is a clique complete to $A_{bf} \cup A_{df}$.

Suppose $u, v \in A_{bf}$ are adjacent; then $G|bcdv \cong \overline{\text{watch}}$. Therefore A_{bf} is a stable set and similarly, so is A_{df} . Now suppose $u \in A_{bf}$ and $v \in A_{df}$ are adjacent. Then $G|bcdv \cong C_5$. Therefore $A_{bf} \cup A_{df}$ is a stable set.

Next, suppose $u, v \in A_{abdef}$ are not adjacent; then $G|bcdv \cong K_{2,3}$. Therefore A_{abdef} is a clique.

Finally, suppose $u \in A_{bf}$ and $v \in A_{abdef}$ are not adjacent. Then $G|abcdev \cong \text{tent}_2$ and u has only one neighbor in $\{a, b, c, d, e, v\}$, which is impossible by (1). Therefore A_{abdef} is complete to A_{bf} and similarly to A_{df} , and this proves (2).

(3) At least one of A_{bf}, A_{df} , and A_{abdef} is empty.

Suppose $u \in A_{bf}$, $v \in A_{df}$, and $w \in A_{abdef}$. From (2), w is complete to $\{u, v\}$ and u is not adjacent to v . It follows that $G|abcdefuvw \cong \overline{F_{22}}$. Therefore at least one of A_{bf}, A_{df} , and A_{abdef} is empty and this proves (3).

If $A_{abdef} = \emptyset$, then it follows from (2) that $G|(A_{bf} \cup A_{df} \cup \{a, e\})$ is a stable set. Also, $G|bcdv$ is antimatched by assumption and $A_{bf} \cup A_{df} \cup \{a, e\}$ and $\{b, c, d, f\}$ are aligned by assumption and definition. Hence, G is doubled.

So we may assume that $A_{abdef} \neq \emptyset$. Then by (3), one of A_{bf} and A_{df} is empty and from symmetry, we may assume A_{df} is empty. Then $G|(A_{bf} \cup \{a, c, d\})$ is semi-matched

with an edge cd , and $G|(A_{abdef} \cup \{b, e, f\})$ is semi-antimatched with a non-edge be . It also follows from assumption and definition that for all $u \in A_{bf} \cup \{a, c, d\}$, u is adjacent to exactly one of b and e and for all $v \in A_{abdef} \cup \{b, e, f\}$, v is adjacent to exactly one of c and d . Hence, G is doubled. This proves 6.3.6. $\square$

We are now ready to prove the main result.

Proof of 6.1.2. The “only if” part is obvious since none of the graphs in $\mathcal{F}$ are doubled. For the “if” part, we may assume G is not almost-split and hence G or $\overline{G}$ contains one of $M_{2,1}$, P_5 , C_6 , domino, tent_1 , and tent_2 as an induced subgraph. But then we are done by 6.3.1, 6.3.2, 6.3.3, 6.3.4, 6.3.5, or 6.3.6 applied to G or $\overline{G}$, keeping in mind that the complement of a doubled graph is doubled. $\square$

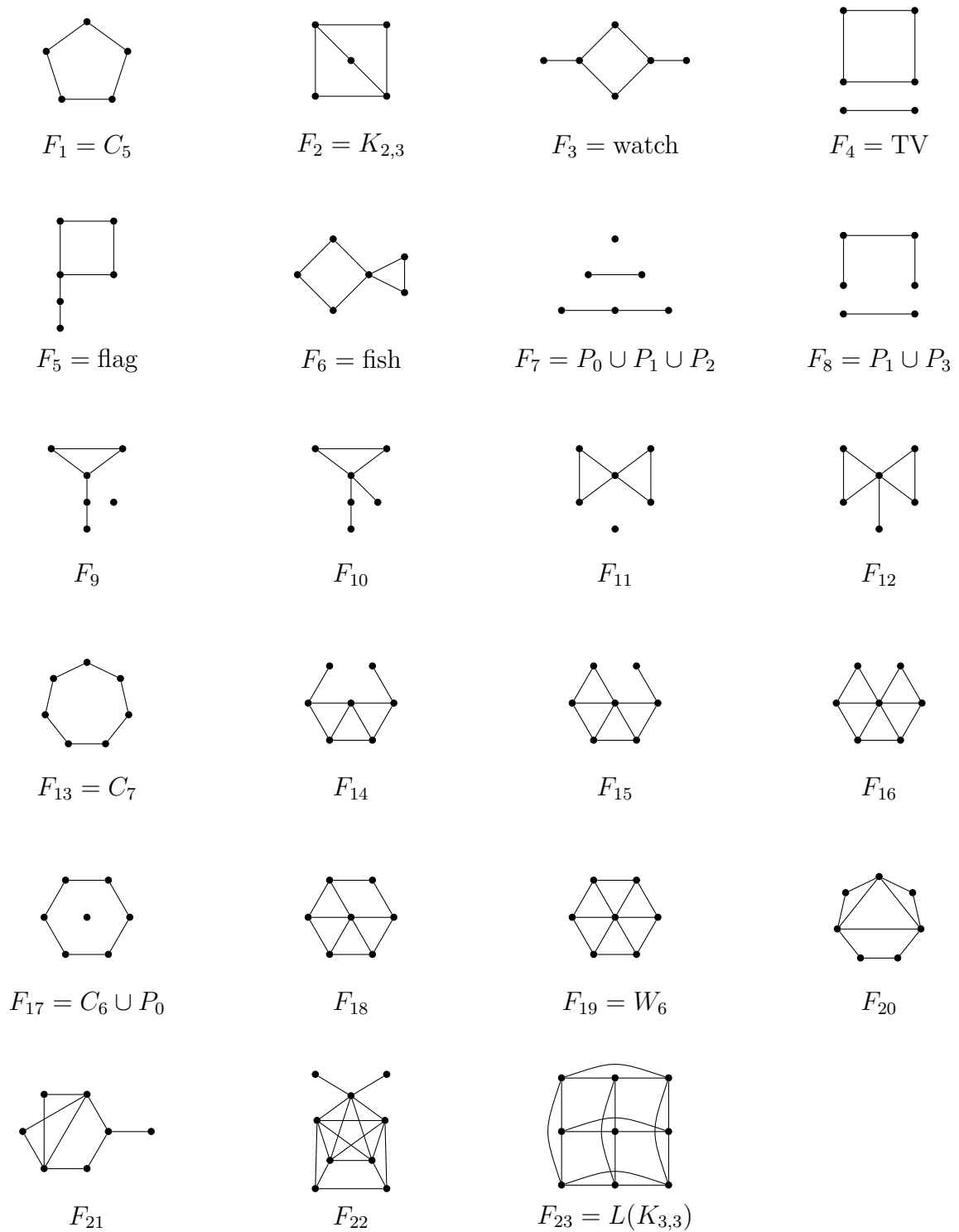


Figure 6.1: The family $\mathcal{F}$: these 23 graphs, and their complements, are the minimal forbidden induced subgraphs for double-split graphs. Only $F_1 = C_5$ and $F_{23} = L(K_{3,3})$ are self-complementary.

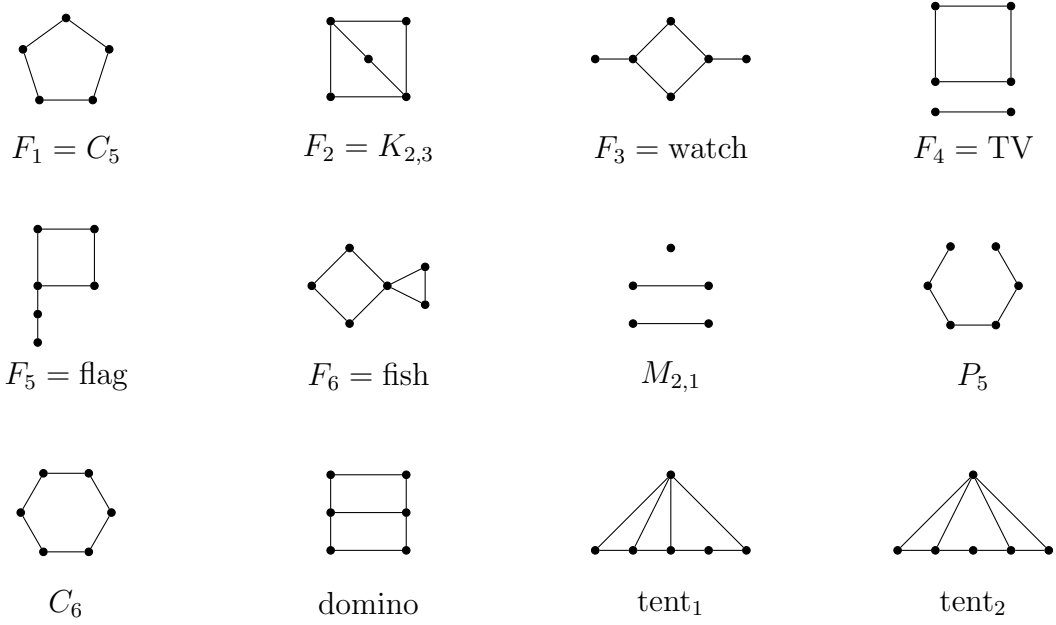


Figure 6.2: The “circus”: these 12 graphs, and their complements, are the minimal forbidden induced subgraphs for almost-split graphs.

Chapter 7

Equations resolving a conjecture of Rado on partition regularity

In 1916, Schur [Sch16] proved that in any coloring of the positive integers with finitely many colors, there is a monochromatic solution to $x + y = z$. In 1927, van der Waerden [vdW27] proved his celebrated theorem that every finite coloring of the positive integers contains arbitrarily long monochromatic arithmetic progressions. In his famous 1933 thesis, Richard Rado [Rad33] generalized these results by classifying the systems of linear equations with monochromatic solutions in every finite coloring. His thesis also contained conjectures regarding equations that *do* have a finite coloring with no monochromatic solutions.

Definition. A linear equation L is k -regular if every k -coloring of the positive integers contains a monochromatic solution to L .

Remark. Some authors require that the values of the variables be distinct in solutions to L . We follow Rado and Guy in not including this extra condition.

Conjecture (Rado, [Rad33] via [Guy04]¹). For every positive integer k , there exists a linear equation that is $(k - 1)$ -regular but not k -regular. In other words, k is the

¹These authors could not verify whether the conjecture is present in Rado's thesis.

least number of colors in a coloring of the positive integers with no monochromatic solution to L .

Fox and Radoičić [FR05] conjectured that the family of equations M_k given by $\sum_{i=0}^{k-2} 2^i x_i = 2^{k-1} x_{k-1}$ has this property. We prove Rado's conjecture by using the related family of equations,

$$\sum_{i=1}^{k-1} \frac{2^i}{2^i - 1} x_i = \left(-1 + \sum_{i=1}^{k-1} \frac{2^i}{2^i - 1} \right) x_0,$$

which we denote by L_k .

Theorem 7.0.7. *The equation L_k is $(k-1)$ -regular but not k -regular.*

Remark. This result and its proof carry over to colorings of the nonzero rationals.

Proof. We first use the power of 2-adic valuations to prove that there exists a k -coloring with no monochromatic solutions to L_k . The idea of using valuations was proposed by Fox and Radoičić for the equation M_k ; later, Alexeev, Fox, and Graham [AFG07a, AFG07b] proved that these colorings were actually minimal if $k \leq 7$, but only conjectured the result in general.

If r is a nonzero rational number, let $\text{ord}_2(r)$ denote the 2-adic valuation of r , that is, the unique integer m such that $r = 2^m \frac{a}{b}$ for odd integers a and b ; also, let $\text{ord}_2(0) = \infty$. Recall that ord_2 satisfies the following two properties:

1. $\text{ord}_2(xy) = \text{ord}_2(x) + \text{ord}_2(y)$,
2. $\text{ord}_2(x + y) \geq \min(\text{ord}_2(x), \text{ord}_2(y))$, with equality if $\text{ord}_2(x)$ does not equal $\text{ord}_2(y)$.

The latter is the (non-Archimedean) 2-adic triangle inequality.

Note that the 2-adic valuation of the coefficient of x_i in L_k is i . This is immediate if $i > 0$ while in the case of $i = 0$, this follows from the 2-adic triangle inequality

because -1 has valuation 0 while the rest of the terms in the summation have positive valuation. We claim this implies that the k -coloring $C_k(r) = \text{ord}_2(r) \bmod k$ has no monochromatic solutions to L_k .

Assume to the contrary that C_k admits a monochromatic solution to L_k . Then the terms of L_k would be a collection of numbers with distinct 2-adic valuations that sum to zero, which is impossible by the 2-adic triangle inequality.

We now show there is no coloring to L_k with fewer than k colors. Indeed, in any coloring with no monochromatic solutions to L_k , the color of x and $2^j x$ must be different if $0 < j < k$. (In the referenced literature, the number 2^j is thus said to be a *forbidden ratio*.) If there were an x so that x and $2^j x$ were the same color, then

$$x_i = \begin{cases} x & \text{if } i = j, \\ 2^j x & \text{if } i \neq j, \end{cases}$$

would be a monochromatic solution to L_k ; all but the j^{th} terms cancel, leaving $\frac{2^j}{2^j-1}x = \left(-1 + \frac{2^j}{2^j-1}\right)2^j x$. This implies that the k numbers $1, 2, 4, \dots, 2^{k-1}$ are colored with distinct colors. □

Chapter 8

Notes

I would like to thank BIRS for organizing the workshop “New trends on structural graph theory”, where the research on double-split graphs (Chapter 6) was initiated.

I would like to thank Owen Biewel for helpful comments on the exposition of Chapter 7.

My coauthors and I would like to thank Swastik Kopparty, Madhu Sudan, Scott Aaronson, Arnab Bhattacharyya, Andy Drucker, Kevin Hughes, Neeraj Kayal, Satya Lokam, Guy Moshkovitz, and Jakob Nordstrom for helpful comments about tensors (Chapter 2).

My coauthors and I would like to thank Peter G. Casazza, Dan Edidin, and Matthew Fickus for insightful discussions on full spark frames (Chapter 4).

My coauthors and I would like to thank Amit Singer for insightful discussions about phaseless reconstruction (Chapter 5).

The works contained within were supported by NSF Grant Nos. DGE-0646086, DMS-0914892, DMS-1008183, DMS-1042701, and DMS-1042701, as well as AFOSR Grant Nos. F1ATA01103J001, F1ATA00183G003, and FA9550-11-1-0245 and the A.B. Krongard Fellowship.

The views expressed throughout this document do not reflect the official policy or

position of the United States Air Force, Department of Defense, or the U.S. Government.

Bibliography

- [ABFM12] Boris Alexeev, Afonso S. Bandeira, Matthew Fickus, and Dustin Mixon, *Phase retrieval with polarization*, Submitted, Oct 2012.
- [ACM12] Boris Alexeev, Jameson Cahill, and Dustin G. Mixon, *Full spark frames*, Journal of Fourier Analysis and Applications **18** (2012), 1167–1194 (English).
- [AFG07a] Boris Alexeev, Jacob Fox, and Ron Graham, *On minimal colorings without monochromatic solutions to a linear equation*, Integers **7** (2007), no. 2, A1, 20 pp. (electronic).
- [AFG07b] ———, *On minimal colorings without monochromatic solutions to a linear equation*, Combinatorial number theory, de Gruyter, Berlin, 2007, pp. 1–22.
- [AFK12] Boris Alexeev, Alexandra Fradkin, and Ilhee Kim, *Forbidden induced subgraphs of double-split graphs*, SIAM Journal on Discrete Mathematics **26** (2012), no. 1, 1–14.
- [AFT11] Boris Alexeev, Michael Forbes, and Jacob Tsimmerman, *Tensor rank: some lower and upper bounds*, IEEE Conference on Computational Complexity, IEEE Computer Society, Feb 2011, pp. 283–291.

- [AFT13] ———, *Tensor rank: some lower and upper bounds*, Journal version with extended results in preparation, Sep 2013.
- [AH59] Michael F. Atiyah and Friedrich Hirzebruch, *Quelques théorèmes de non-plongement pour les variétés différentiables*, Bull. Soc. Math. France **87** (1959), 383–396.
- [Alo86] N. Alon, *Eigenvalues and expanders*, Combinatorica **6** (1986), no. 2, 83–96, Theory of computing (Singer Island, Fla., 1984).
- [AM85] N. Alon and V. D. Milman, λ_1 , *isoperimetric inequalities for graphs, and superconcentrators*, J. Combin. Theory Ser. B **38** (1985), no. 1, 73–88.
- [Art91] Michael Artin, *Algebra*, Prentice Hall Inc., Englewood Cliffs, NJ, 1991.
- [AT10] Boris Alexeev and Jacob Tsimerman, *Equations resolving a conjecture of Rado on partition regularity*, J. Combin. Theory Ser. A **117** (2010), no. 7, 1008–1010.
- [AW10] Boris Alexeev and Rachel Ward, *On the complexity of Mumford-Shah type regularization, viewed as a relaxed sparsity constraint*, IEEE Trans. Image Process. **19** (2010), no. 10, 2787–2789.
- [BBCE07] Radu Balan, Bernhard G. Bodmann, Peter G. Casazza, and Dan Edidin, *Fast algorithms for signal reconstruction without phase*, Wavelets XII, vol. 6701, 2007, pp. 67011L–67011L–9.
- [BBCE09] ———, *Painless reconstruction from magnitudes of frame coefficients*, J. Fourier Anal. Appl. **15** (2009), no. 4, 488–501.
- [BCE06] Radu Balan, Pete Casazza, and Dan Edidin, *On signal reconstruction without phase*, Appl. Comput. Harmon. Anal. **20** (2006), no. 3, 345–356.

- [BCM12] Waheed U. Bajwa, Robert Calderbank, and Dustin G. Mixon, *Two are better than one: fundamental parameters of frame coherence*, Appl. Comput. Harmon. Anal. **33** (2012), no. 1, 58–78.
- [BD80] Mark R. Brown and David P. Dobkin, *An improved lower bound on polynomial multiplication*, IEEE Trans. Comput. **29** (1980), no. 5, 337–340.
- [BDP⁺07] Oliver Bunk, Ana Diaz, Franz Pfeiffer, Christian David, Bernd Schmitt, Dillip K. Satapathy, and J. Friso van der Veen, *Diffraction imaging for periodic samples: retrieving one-dimensional concentration profiles across microfluidic channels*, Acta Crystallographica Section A **63** (2007), no. 4, 306–314.
- [Bes89] Julian E. Besag, *Digital image processing: Toward Bayesian image analysis*, J. Appl. Statist. **16** (1989), no. 3, 395–407.
- [Blä99] Markus Bläser, *A $\frac{5}{2}n^2$ -lower bound for the rank of $n \times n$ -matrix multiplication over arbitrary fields*, 40th Annual Symposium on Foundations of Computer Science (New York, 1999), IEEE Computer Soc., Los Alamitos, CA, 1999, pp. 45–50.
- [BS83] Walter Baur and Volker Strassen, *The complexity of partial derivatives*, Theoret. Comput. Sci. **22** (1983), no. 3, 317–330.
- [BSS12] A. S. Bandeira, A. Singer, and D. A. Spielman, *A Cheeger Inequality for the Graph Connection Laplacian*, preprint, April 2012.
- [BZ87] Andrew Blake and Andrew Zisserman, *Visual reconstruction*, MIT Press Series in Artificial Intelligence, MIT Press, Cambridge, MA, 1987.

- [CCH11] Jameson Cahill, Peter G. Casazza, and Andreas Heinecke, *A notion of redundancy for infinite frames*, Proc. Sampl. Theory Appl (2011), 4pp.
- [CESV13] Emmanuel J. Candès, Yonina C. Eldar, Thomas Strohmer, and Vladislav Voroninski, *Phase retrieval via matrix completion*, SIAM J. Imaging Sci. **6** (2013), no. 1, 199–225.
- [Cha95] Antonin Chambolle, *Image segmentation by variational methods: Mumford and Shah functional and the discrete approximations*, SIAM J. Appl. Math. **55** (1995), no. 3, 827–863.
- [Che70] Jeff Cheeger, *A lower bound for the smallest eigenvalue of the Laplacian*, Problems in analysis (Papers dedicated to Salomon Bochner, 1969), Princeton Univ. Press, Princeton, N. J., 1970, pp. 195–199.
- [CHKK11] Peter G. Casazza, Andreas Heinecke, Felix Krahmer, and Gitta Kutyniok, *Optimally sparse frames*, IEEE Trans. Inform. Theory **57** (2011), no. 11, 7279–7287.
- [Chu10] Fan Chung, *Four proofs for the Cheeger inequality and graph partition algorithms*, Fourth International Congress of Chinese Mathematicians, AMS/IP Stud. Adv. Math., vol. 48, Amer. Math. Soc., Providence, RI, 2010, pp. 331–349.
- [CRST06] Maria Chudnovsky, Neil Robertson, Paul Seymour, and Robin Thomas, *The strong perfect graph theorem*, Ann. of Math. (2) **164** (2006), no. 1, 51–229.
- [CSV11] E. J. Candès, T. Strohmer, and V. Voroninski, *PhaseLift: Exact and Stable Signal Recovery from Magnitude Measurements via Convex Programming*, preprint, September 2011.

- [CT05] Emmanuel J. Candes and Terence Tao, *Decoding by linear programming*, IEEE Trans. Inform. Theory **51** (2005), no. 12, 4203–4215.
- [CW08] E. J. Candes and M. B. Wakin, *An introduction to compressive sampling*, Signal Processing Magazine, IEEE **25** (2008), no. 2, 21–30.
- [DDEK12] Mark A. Davenport, Marco F. Duarte, Yonina C. Eldar, and Gitta Kutyniok, *Introduction to compressed sensing*, Compressed sensing, Cambridge Univ. Press, Cambridge, 2012, pp. 1–64.
- [DE03] David L. Donoho and Michael Elad, *Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization*, Proc. Natl. Acad. Sci. USA **100** (2003), no. 5, 2197–2202 (electronic).
- [DF87] J. C. Dainty and J. R. Fienup, *Phase retrieval and image reconstruction for astronomy*, Image Recovery: Theory and Application, ed. by H. Stark, Academic Press, San Diego (1987), 231–275.
- [DH12] L. Demanet and P. Hand, *Stable optimizationless recovery from phaseless linear measurements*, preprint, August 2012.
- [DS06] Ken Dykema and Nate Strawn, *Manifold structure of spaces of spherical tight frames*, Int. J. Pure Appl. Math. **28** (2006), no. 2, 217–256.
- [DVB08] Steven Delvaux and Marc Van Barel, *Rank-deficient submatrices of Fourier matrices*, Linear Algebra Appl. **429** (2008), no. 7, 1587–1605.
- [EI76] R. J. Evans and I. M. Isaacs, *Generalized Vandermonde determinants and roots of unity of prime order*, Proc. Amer. Math. Soc. **58** (1976), 51–54. MR 0412205 (54 #332)
- [FH77] Stéphane Foldes and Peter L. Hammer, *Split graphs*, Proceedings of the Eighth Southeastern Conference on Combinatorics, Graph Theory and

- Computing (Louisiana State Univ., Baton Rouge, La., 1977) (Winnipeg, Man.), *Utilitas Math.*, 1977, pp. 311–315. *Congressus Numerantium*, No. XIX.
- [Fie82] J. R. Fienup, *Phase retrieval algorithms: a comparison*, *Appl. Opt.* **21** (1982), no. 15, 2758–2769.
- [FM11] M. Fickus and D. G. Mixon, *Deterministic matrices with the restricted isometry property*, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 8138, September 2011.
- [FM12] Matthew Fickus and Dustin G. Mixon, *Numerically erasure-robust frames*, *Linear Algebra Appl.* **437** (2012), no. 6, 1394–1407.
- [FMT12] Matthew Fickus, Dustin G. Mixon, and Janet C. Tremain, *Steiner equiangular tight frames*, *Linear Algebra Appl.* **436** (2012), no. 5, 1014–1027.
- [FR05] Jacob Fox and Rados Radoičić, *The axiom of choice and the degree of regularity of equations over the reals*, unpublished preprint, Dec 2005.
- [Fri08] Joel Friedman, *A proof of Alon’s second eigenvalue conjecture and related problems*, *Mem. Amer. Math. Soc.* **195** (2008), no. 910, viii+100.
- [FW10] Massimo Fornasier and Rachel Ward, *Iterative thresholding meets free-discontinuity problems*, *Found. Comput. Math.* **10** (2010), no. 5, 527–567.
- [GG84] Stuart Geman and Donald Geman, *Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images*, *IEEE Trans. Pattern Anal. Mach. Intell.* **6** (1984), no. 6, 721–741.

- [GJLM84] Daniel W. Griffin, Jae, S. Lim, and Senior Member, *Signal estimation from modified short-time fourier transform*, IEEE Trans. Acoustics, Speech and Sig. Proc (1984), 236–243.
- [Goy98] Vivek K Goyal, *Beyond traditional transform coding*, Ph.D. thesis, University California, Berkeley, 1998.
- [GP97] Shuhong Gao and Daniel Panario, *Tests and constructions of irreducible polynomials over finite fields*, Foundations of computational mathematics (Rio de Janeiro, 1997), Springer, Berlin, 1997, pp. 346–361.
- [GS72] R. Gerchberg and W. Saxton, *A practical algorithm for the determination of phase from image and diffraction plane pictures*, Optik (1972), 237–246.
- [Guy04] Richard K. Guy, *Unsolved problems in number theory*, third ed., Problem Books in Mathematics, Springer-Verlag, New York, 2004.
- [Hal35] P. Hall, *On representatives of subsets*, Journal of the London Mathematical Society **s1-10** (1935), no. 1, 26–30.
- [Har93] Robert W. Harrison, *Phase problem in crystallography*, J. Opt. Soc. Am. A **10** (1993), no. 5, 1046–1055.
- [Hås89] Johan Håstad, *Tensor rank is NP-complete*, Automata, languages and programming (Stresa, 1989), Lecture Notes in Comput. Sci., vol. 372, Springer, Berlin, 1989, pp. 451–460.
- [Hås90] ———, *Tensor rank is NP-complete*, J. Algorithms **11** (1990), no. 4, 644–654.

- [HB05] P. Harsha and A. Barth, *Lecture 5: Derandomization (part ii)*, lecture notes available at <http://www.tcs.tifr.res.in/~prahladh/teaching/05spring/lectures/lec5.pdf>, May 2005.
- [HK71] J. E. Hopcroft and L. R. Kerr, *On minimizing the number of multiplications necessary for matrix multiplication*, SIAM J. Appl. Math. **20** (1971), 30–36.
- [HMW13] Teiko Heinosaari, Luca Mazzarella, and Michael M. Wolf, *Quantum tomography under prior information*, Communications in Mathematical Physics **318** (2013), no. 2, 355–374 (English).
- [Hoe63] Wassily Hoeffding, *Probability inequalities for sums of bounded random variables*, J. Amer. Statist. Assoc. **58** (1963), 13–30.
- [HP04] Roderick B. Holmes and Vern I. Paulsen, *Optimal frames for erasures*, Linear Algebra Appl. **377** (2004), 31–51.
- [Ja’79] Joseph Ja’Ja’, *Optimal evaluation of pairs of bilinear forms*, SIAM J. Comput. **8** (1979), no. 3, 443–462.
- [JPS07] D. Jungnickel, A. Pott, and K.W. Smith, *Difference sets.*, Handbook of Combinatorial Designs (Charles J. Colbourn and Jeffrey H. Dinitz, eds.), Discrete Mathematics and its Applications (Boca Raton), Chapman & Hall/CRC, Boca Raton, FL, second ed., 2007, pp. 419–435.
- [JW90] Fure-Ching Jeng and John W. Woods, *Simulated annealing in compound Gaussian random fields*, IEEE Trans. Inform. Theory **36** (1990), no. 1, 94–107.
- [Kal85] K. A. Kalorkoti, *A lower bound for the formula size of rational functions*, SIAM J. Comput. **14** (1985), no. 3, 678–687.

- [Kam05] Michael Kaminski, *A lower bound on the complexity of polynomial multiplication over finite fields*, SIAM J. Comput. **34** (2005), no. 4, 960–992 (electronic).
- [Kar72] Richard M. Karp, *Reducibility among combinatorial problems*, Complexity of computer computations (Proc. Sympos., IBM Thomas J. Watson Res. Center, Yorktown Heights, N.Y., 1972), Plenum, New York, 1972, pp. 85–103.
- [Lan13] J. M. Landsberg, *Explicit tensors of border rank at least $2n-1$* , preprint, August 2013.
- [Li11] Yang D Li, *Applications of monotone rank to complexity theory*, preprint, 2011.
- [LM00] B. Laurent and P. Massart, *Adaptive estimation of a quadratic functional by model selection*, Ann. Statist. **28** (2000), no. 5, 1302–1338.
- [LPS88] A. Lubotzky, R. Phillips, and P. Sarnak, *Ramanujan graphs*, Combinatorica **8** (1988), no. 3, 261–277.
- [Mar09] Dániel Marx, *A parameterized view on matroid optimization problems*, Theoret. Comput. Sci. **410** (2009), no. 44, 4471–4479.
- [McC83] Seth Thomas McCormick, *Combinatorial approach to some sparse-matrix problems*, Tech. report, Stanford Univ., CA (USA). Dept. of Operations Research, 1983.
- [Mil90] R. P. Millane, *Phase retrieval in crystallography and optics*, J. Opt. Soc. Am. A **7** (1990), no. 3, 394–411.

- [MISE08] J. Miao, T. Ishikawa, Q. Shen, and T. Earnest, *Extending X-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes*, Annu Rev Phys Chem **59** (2008), 387–410.
- [MQKF11] D.G. Mixon, C. Quinn, N. Kiyavash, and M. Fickus, *Equiangular tight frame fingerprinting codes*, Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, 2011, pp. 1856–1859.
- [Nik00] Mila Nikolova, *Thresholding implied by truncated quadratic regularization*, IEEE Trans. Signal Process. **48** (2000), no. 12, 3437–3450.
- [NM82] S. Nakamura and G.M. Masson, *Lower bounds on crosspoints in concentrators*, Computers, IEEE Transactions on **C-31** (1982), no. 12, 1173–1179.
- [PK05] M. Puschel and J. Kovacevic, *Real, tight frames with maximal robustness to erasures*, Data Compression Conference, 2005. Proceedings. DCC 2005, 2005, pp. 63–72.
- [PW70] M. J. Piff and D. J. A. Welsh, *On the vector representation of matroids*, J. London Math. Soc. (2) **2** (1970), 284–288.
- [Rad33] Richard Rado, *Studien zur Kombinatorik*, Math. Z. **36** (1933), no. 1, 424–470.
- [Rap02] T. Rapcsák, *On minimization on Stiefel manifolds*, European J. Oper. Res. **143** (2002), no. 2, 365–376, Interior point methods (Budapest, 2000).
- [Raz10] Ran Raz, *Tensor-rank and lower bounds for arithmetic formulas*, STOC’10—Proceedings of the 2010 ACM International Symposium on Theory of Computing, ACM, New York, 2010, pp. 659–666.

- [RBKSC04] Joseph M. Renes, Robin Blume-Kohout, A. J. Scott, and Carlton M. Caves, *Symmetric informationally complete quantum measurements*, J. Math. Phys. **45** (2004), no. 6, 2171–2180.
- [Ren07] Joseph M. Renes, *Equiangular tight frames from Paley tournaments*, Linear Algebra Appl. **426** (2007), no. 2-3, 497–501.
- [Ron07] Luca Rondi, *A variational approach to the reconstruction of cracks by boundary measurements*, J. Math. Pures Appl. (9) **87** (2007), no. 3, 324–342.
- [Ron08] ———, *Reconstruction in the inverse crack problem by variational methods*, European J. Appl. Math. **19** (2008), no. 6, 635–660.
- [RSA⁺11] O. Raz, O. Schwartz, D. Austin, A. S. Wyatt, A. Schiavi, O. Smirnova, B. Nadler, I. A. Walmsley, D. Oron, and N. Dudovich, *Vectorial phase retrieval for linear characterization of attosecond pulses*, Phys. Rev. Lett. **107** (2011), 133902.
- [RV08] Mark Rudelson and Roman Vershynin, *On sparse reconstruction from Fourier and Gaussian measurements*, Comm. Pure Appl. Math. **61** (2008), no. 8, 1025–1045.
- [SC91] H. Sahinoglou and S.D. Cabrera, *On phase retrieval of finite-length sequences using the initial time sample*, IEEE Transactions on Circuits and Systems **38** (1991), no. 8, 954–958.
- [Sch16] Issai Schur, *Über die Kongruenz $x^m + y^m \equiv z^m \pmod{p}$* , Jahresber Deutsch. Math. Verein. **25** (1916), 114–117.

- [Ser77] Jean-Pierre Serre, *Linear representations of finite groups*, Springer-Verlag, New York, 1977, Translated from the second French edition by Leonard L. Scott, Graduate Texts in Mathematics, Vol. 42.
- [SH03] Thomas Strohmer and Robert W. Heath, Jr., *Grassmannian frames with applications to coding and communication*, Appl. Comput. Harmon. Anal. **14** (2003), no. 3, 257–275.
- [Sho90] Victor Shoup, *New algorithms for finding irreducible polynomials over finite fields*, Math. Comp. **54** (1990), no. 189, 435–447.
- [Shp03] Amir Shpilka, *Lower bounds for matrix product*, SIAM J. Comput. **32** (2003), no. 5, 1185–1200 (electronic).
- [Sin11] A. Singer, *Angular synchronization by eigenvectors and semidefinite programming*, Appl. Comput. Harmon. Anal. **30** (2011), no. 1, 20–36.
- [SL96] P. Stevenhagen and H. W. Lenstra, Jr., *Chebotarëv and his density theorem*, Math. Intelligencer **18** (1996), no. 2, 26–37.
- [SMS09] Toshio Sumi, Mitsuhiro Miyazaki, and Toshio Sakata, *Rank of 3-tensors with 2 slices and Kronecker canonical forms*, Linear Algebra Appl. **431** (2009), no. 10, 1858–1868.
- [Str73] Volker Strassen, *Vermeidung von Divisionen*, J. Reine Angew. Math. **264** (1973), 184–202.
- [Str75] ———, *Die Berechnungskomplexität der symbolischen Differentiation von Interpolationspolynomen*, Theor. Comput. Sci. **1** (1975), no. 1, 21–25.
- [Str08] Thomas Strohmer, *A note on equiangular tight frames*, Linear Algebra Appl. **429** (2008), no. 1, 326–330.

- [Str11] Nate Strawn, *Finite frame varieties: nonsingular points, tangent spaces, and explicit local parameterizations*, J. Fourier Anal. Appl. **17** (2011), no. 5, 821–853.
- [SW01] Amir Shpilka and Avi Wigderson, *Depth-3 arithmetic circuits over fields of characteristic zero*, Comput. Complexity **10** (2001), no. 1, 1–27.
- [Tao05] Terence Tao, *An uncertainty principle for cyclic groups of prime order*, Math. Res. Lett. **12** (2005), no. 1, 121–127.
- [Tro08] Joel A. Tropp, *On the conditioning of random subdictionaries*, Appl. Comput. Harmon. Anal. **25** (2008), no. 1, 1–24.
- [Val80] L. G. Valiant, *Negation can be exponentially powerful*, Theoret. Comput. Sci. **12** (1980), no. 3, 303–314.
- [vdW27] Bartel Leendert van der Waerden, *Beweis einer Baudet’schen Vermutung*, Nieuw Arch. Wiskunde **15** (1927), 212–216.
- [Ver10] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, preprint, November 2010.
- [VK85] A. M. Vershik and S. V. Kerov, *Asymptotic behavior of the maximum and generic dimensions of irreducible representations of the symmetric group*, Funktsional. Anal. i Prilozhen. **19** (1985), no. 1, 25–36, 96.
- [vL99] J. H. van Lint, *Introduction to coding theory*, third ed., Graduate Texts in Mathematics, vol. 86, Springer-Verlag, Berlin, 1999.
- [vLW01] J. H. van Lint and R. M. Wilson, *A course in combinatorics*, second ed., Cambridge University Press, Cambridge, 2001.
- [VV86] L. G. Valiant and V. V. Vazirani, *NP is as easy as detecting unique solutions*, Theoret. Comput. Sci. **47** (1986), no. 1, 85–93.

- [vzG88] Joachim von zur Gathen, *Algebraic complexity theory*, Annual review of computer science, Vol. 3, Annual Reviews, Palo Alto, CA, 1988, pp. 317–347.
- [Wal63] A. Walther, *The Question of Phase Retrieval in Optics*, Optica Acta **10** (1963), 41–49.
- [WdM12] I. Waldspurger, A. d’Aspremont, and S. Mallat, *Phase Recovery, MaxCut and Complex Semidefinite Programming*, preprint, June 2012.
- [XZG05] Pengfei Xia, Shengli Zhou, and Georgios B. Giannakis, *Achieving the Welch bound with difference sets*, IEEE Trans. Inform. Theory **51** (2005), no. 5, 1900–1907.
- [Zau99] G. Zauner, *Quantendesigns: Grundzüge einer nichtkommutativen designtheorie*, Ph.D. thesis, University of Vienna, 1999.
- [ZZ05] Igor E. Zverovich and Vadim E. Zverovich, *Basic perfect graphs and their extensions*, Discrete Math. **293** (2005), no. 1-3, 291–311.