Fourier Analysis on the Hypercube, the Coefficient Problem, and Applications

by

Ganesh Ajjanagadde

S.B., Massachusetts Institute of Technology (2015)
M.Eng., Massachusetts Institute of Technology (2016)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

Author . . . .
Department of Electrical Engineering and Computer Science
April 28, 2020

Certified by . . . .
Gregory Wornell
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Certified by . . . .
Henry Cohn
Senior Principal Researcher, Microsoft Research New England
Thesis Supervisor

Accepted by . . . .
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Fourier Analysis on the Hypercube, the Coefficient Problem, and Applications

by

Ganesh Ajjanagadde

Submitted to the Department of Electrical Engineering and Computer Science
on April 28, 2020, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science

Abstract

In this dissertation, we primarily study some problems that revolve around Fourier analysis. More specifically we focus on the magnitudes of the frequency components. Firstly, we perform a study on the hypercube. It is well known that the Delsarte linear programming bounds provide rich information on the magnitudes of the Fourier coefficients, grouped by Hamming weight. Classically, such information is primarily used to attack coding problems, where the objective is to maximize cardinality of a subset of a metric space subject to a minimum distance constraint. Here, we use it to study anticoding problems, where the objective is to maximize cardinality of a subset of a metric space subject to a maximum distance (diameter) constraint. One motivation for such study is the problem of finding memories that are cheap to update, where the cost of an update is a function of the distance in the metric space. Such a view naturally supports the study of different cost functions going beyond hard diameter constraints. We work accordingly with different cost functions, with a particular emphasis on completely monotonic functions. Our emphasis is on the phenomenon of "universal optimality", where the same subset (anticode) simultaneously optimizes a wide range of natural cost functions. Among other things, our work here gives some answers to a question in computer science, namely finding Boolean functions with maximal noise stability subjected to an expected value constraint.

Secondly, we work with Fourier analysis on the integers modulo a number by drawing upon Nazarov's general solution to the "coefficient problem". Roughly speaking, the coefficient problem asks one to construct time domain signals with prescribed magnitudes of frequency components, subject to certain natural constraints on the signal. In particular, Nazarov's solution works with l_p constraints in time. This solution to the coefficient problem allows us to give an essentially complete resolution to the mathematical problem of designing optimal coded apertures that arises in computational imaging. However, the resolution we provide is for an l_∞ constraint on the aperture, corresponding to partial occlusion. We believe it is important to also examine a binary valued ($\{0, 1\}$) constraint on the aperture as one does not need to synthesize partial occluders for such apertures. We therefore provide some preliminary results as well as directions for future research.

Finally, inspired by the recent breakthroughs in understanding the $d = 8, 24$ cases of sphere packing and universal optimality in $\mathbb{R}^d$, we attempt to show that the associated lattices (E_8 and the Leech lattice for $d = 8, 24$ respectively) are also optimal for the problem of vector quantization in the sense of minimizing mean squared error. Accordingly, we develop a dispersion and anticoding based approach to lower bounds on the mean squared error. We also generalize Tóth's method, which shows optimality of the hexagonal lattice quantizer for $d = 2$, to arbitrary d. To the best of our knowledge, these methods give the first rigorous improved lower bounds for the mean squared error for all large enough d since the work of Zador over 50 years ago.

Thesis Supervisor: Gregory Wornell
Title: Professor of Electrical Engineering and Computer Science

Thesis Supervisor: Henry Cohn
Title: Senior Principal Researcher, Microsoft Research New England

Acknowledgments

It is a Herculean task to express my gratitude to all who have contributed directly and indirectly to my adventures in graduate school and in this dissertation. I can only offer a few succinct remarks here. For those of you not explicitly mentioned here: first off, I am confident that you are already aware of my deep respect for you, and that you would not derive much meaning or pleasure from an explicit acknowledgement here. Secondly, I am grateful for the very fact that you are examining this dissertation. I hope that you find it at best illuminating, and at worst somewhat entertaining. Keeping this in my mind, I now turn to some of the most important people, hoping that you understand how much you have meant to me.

First, I thank my parents, Venkataramana and Vijaya Ajjanagadde. To put it simply: I am who I am because of you; any merits I might have are due to you, and any demerits are mine alone.

Second, I thank my thesis advisers, Profs. Gregory Wornell and Henry Cohn. Both of you have not only been extremely patient, supportive, infectiously optimistic, and insightful with respect to research, but have also extended it to my development as a person. It was an honor to be your student.

Third, I thank Prof. Yury Polyanskiy, my thesis committee member. It is pretty safe to say that I was forged as a researcher in his smithy as an undergraduate. I am also grateful to him for exposing me to the wonders of the Russian intelligentsia.

Fourth, I acknowledge several professors with whom I have had stimulating interactions over the past eight years here at MIT: Profs. Guy Bresler, Polina Golland, Alexandre Megretski, Elchanan Mossel, John Tsitsiklis, George Verghese, Alan Willsky, and Lizhong Zheng.

Fifth, I acknowledge the friends who have served as a bedrock during my time here at MIT: Mohamed AlHajri, Nirav Bhan, Kishor Bhat, Austin Collins, Matthew de Courcy-Ireland, Igor Kadota, Pranav Kaundinya, Joshua Ka-Wing Lee, Eren Kizildag, Suhas Kowshik, Fabián Kozynski, Pavitra Krishnaswamy, Ashwin Kumar, Tarek Lahlou, Bhavesh Lalwani, Anuran Makur, Dheeraj Nagaraj, Deepak Narayanan,

James Noraky, Or Ordentlich, Rishi Patel, David Qiu, Govind Ramnarayan, Ankit Rawat, Arman Rezaee, Hajir Roozbehani, Tuhin Sarkar, Anjan Soumyanarayanan, Harihar Subramanyam, James Thomas, Christos Thrampoulidis, Aditya Venkatramani, and Adam Yedidia.

Sixth, I acknowledge the outstanding labmates (past and present) who have made the Signals, Information, and Algorithms Lab a special place: Toros Arikan, Yuheng Bu, Qing He, Tejas Jayashankar, Gauri Joshi, Gary Lee, Emin Martinian, Safa Medin, Gal Shulkind, Atulya Yellepeddi, and Xuhong Zhang. I also acknowledge the always friendly and helpful administrative assistant Tricia O'Donnell.

Finally, I thank my extended family. They have been a very deep foundation during the times when I could not see the light at the end of the tunnel. In particular, I dedicate this thesis to my grandmothers Parvathi Ajjanagadde and Shankari Bhat, who instilled a deep love for education and knowledge in their descendants in spite of very difficult circumstances.

The material in this dissertation is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. CCF-1816209. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Much of the work presented in Chapter 4 was performed while I was at Microsoft Research New England, whose support I gratefully acknowledge.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> Regarding the researches of d'Alembert and Euler could one not add that if they knew this expansion, they made but a very imperfect use of it. They were both persuaded that an arbitrary and discontinuous function could never be resolved in series of this kind, and it does not even seem that anyone had developed a constant in cosines of multiple arcs, the first problem which I had to solve in the theory of heat.
>
> *Joseph Fourier,*1808-9

1.1 Motivation

In science and engineering, the modern problems we face often have a rich history lying underneath their surface. Understanding this history is often crucial in the resolution of these problems. This dissertation attempts to defend this point of view through a set of case studies, all revolving around the topic of Fourier analysis.

Fourier analysis may be attributed to the work of Joseph Fourier on the theory of heat transfer [48]. Roughly speaking, Fourier analysis allows one to study the way a general function can be represented as a linear combination of simpler trigonometric functions (sinusoids). The viewpoint of a function as a combination of sinusoids is certainly a very useful one, and seems to be how Fourier himself envisioned it.

Over time, however, as Fourier's ideas were cultivated and their impact realized, more sophisticated points of view emerged. For example, the trigonometric functions can be replaced by complex exponentials, and the fundamental role of the complex exponentials comes from the fact that they are simultaneous eigenfunctions for the translation group generated from the basic translation: $(Tf)(x) \triangleq f(x+1)$.

Such a viewpoint may be developed further into a theory of the Fourier transform on groups. In the abelian case (including Hamming space $(\mathbb{Z}/q\mathbb{Z})^d$), the theory is simpler as the Fourier transform remains scalar valued, and the fundamental dual objects are *characters*. Finiteness of the group also helps keep things simple. Details at a level suitable for both engineers and mathematicians may be found in e.g. the wonderful book of Terras [112], or the article of Forney [47]. We shall review this material in the context of finite abelian groups, such as Hamming space $(\mathbb{Z}/q\mathbb{Z})^d$, with minimal prerequisites. The elementary approach for finite abelian groups has the advantage of being accessible to more people, though to appreciate the generality and richness of these ideas and plumb deeper one can study the representation theory of finite groups (e.g. [103, 72]) and compact/locally compact groups (e.g. [97]).

We now briefly describe the contents of this dissertation and how they relate to Fourier analysis on three spaces, namely $(\mathbb{Z}/q\mathbb{Z})^d, \mathbb{Z}/n\mathbb{Z}, \mathbb{R}^d$.

1.1.1 Hamming space

The primary motivation for our study on Hamming space in Chapter 2 is understanding "anticodes" better. Here, the classical question is to maximize the size of a subset of $\{0,1\}^n$ subject to an *upper bound* on distances between pairs of points. We call this problem a *isodiametric problem*. The reason we call such subsets "anticodes" is that one can replace the *upper bound* with *lower bound* to yield the classical central question of coding theory. It is somewhat remarkable that in spite of the coding theory question remaining unresolved, Ahlswede and Khachatrian [7, Diametric Theorem] obtained a complete resolution to the anticoding question in the above isodiametric sense.

Here, we generalize the definition of optimality from the isodiametric sense to that of optimizing a two point potential function subject to a cardinality constraint.

We employ the classical Delsarte linear programming (LP) bounds, which are intimately connected to Fourier analysis. The LP bounds turn out to yield sharp answers for some special values of the cardinality. We give an application to a problem in theoretical computer science, namely that of finding a Boolean function $f : \{-1, 1\}^n \to \{-1, 1\}$ which maximizes noise stability subject to an expected value constraint. Previously, sharp answers were known only for $\mathbb{E}(f) = 0$, where the answer is a dictator function $f(x^n) = x_1$ irrespective of the value of noise. Here, we resolve $\mathbb{E}(f) = \pm 1/2$, where an answer is given by $f(x^n) = x_1 \wedge x_2$ and $f(x^n) = \overline{x_1 \wedge x_2}$, irrespective of the value of noise. We generalize such results to a non-binary setting, namely $f : \{0, 1, \ldots, q-1\}^n \to \{-1, 1\}$. We also exhibit a stacking construction of anticodes, and utilize it to prove that the set of universal optima for noise stability across the noise level is a sparse set.

1.1.2 Discrete Fourier transforms

The primary motivation for our study of the discrete Fourier transform in Chapter 3 is computational imaging, specifically the problem of designing good coded aperture systems. Roughly speaking, coded aperture imaging systems consist of a perforated plate placed before the imaging plane, and an associated computational inversion procedure to recover the scene of interest from the image formed by a superposition of shifted copies of the scene (a convolution) through the various perforations. The basic design problem is to come up with a good perforation pattern.

In [8], we characterize the fundamental limits of coded aperture imaging systems up to universal constants by drawing upon a theorem of Nazarov regarding Fourier transforms. The theorem itself is more general, and we will elaborate on this. Our work is performed under a simple propagation and sensor model that accounts for thermal and shot noise, scene correlation, and exposure time. Focusing on mean square error as a measure of linear reconstruction quality, we show that appropriate application of a theorem of Nazarov leads to essentially optimal coded apertures, up to a constant multiplicative factor in exposure time. Additionally, we develop a heuristically efficient algorithm to generate such patterns that explicitly takes into account scene correlations. This algorithm finds apertures that correspond to local optima of a certain potential on the hypercube, yet are guaranteed to be tight. Finally, for i.i.d. scenes, we show improvements upon prior work by using spectrally flat sequences with bias. The development primarily focuses on one dimensional apertures for conceptual clarity; the natural generalizations to 2D are also discussed.

1.1.3 Euclidean space

Fourier analysis played a crucial role in the resolution of the sphere packing problem (the coding problem for Euclidean space) and associated universal optimality phenomena for $\mathbb{R}^d, d = 8, 24$ [120], [27], [28]. We agree with the experts and believe that the associated universally optimal structures, namely the lattices E_8 for $d = 8$ and the Leech lattice for $d = 24$ are also optimal for the quantization problem in the sense of minimizing mean squared error in the so-called "high-resolution limit", and refer the impatient reader to Conjecture 6 for a precise statement.

We still believe that Fourier analysis will play a role in resolving Conjecture 6, but are currently unable to shed any light on such an idea. Instead, we develop alternative methods that are still capable of yielding improved lower bounds on the mean squared error for lattice as well as non-lattice quantizers. These bounds represent the first rigorous improvement over Zador's sphere bound [128], [129], though Conway and Sloane have a conjectured bound [31]. We obtain distinct lower bounds for lattice quantizers versus general quantizers, with stronger results for lattice quantizers. The results for lattices are numerically verifiable albeit conjectural. In either setting, our bounds are not as strong as the one conjectured by Conway and Sloane. At a high level, our approach may be viewed as a generalization of the work of Tóth/Newman [116], [88] from dimension $d = 2$ to larger d. We achieve this by utilizing upper bounds on face counts of Voronoï cells from Minkowski/Voronoï [84], [121] in the lattice case, and considerations of dispersion and upper bounds on sphere packing density in the general case. One route to the sphere packing density is through LP bounds. This second "dispersion method" thus indirectly relies on Fourier analysis, through the links between the LP bounds and Fourier analysis that we first alluded to above with

Hamming space and that we will describe in greater detail in Chapter 2. However, there are other approaches to nontrivial upper bounds on packing density, such as Rankin's method [94], which extended earlier work of Blichfeldt [20]. As such, we do not recommend reading too much into this link with Fourier analysis.

1.1.4 General remarks

We shall develop these ideas in a self-contained manner in as elementary a fashion as possible. In particular, we provide proofs of facts that may be found in the original sources or other expositions either explicitly or implicitly, unless they are well known to a general audience of scientists/engineers, or take us too far away from our main thread. As a concrete example, we assume Cauchy-Schwarz is a well-known inequality, but that the discrete Fourier transform of quadratic residue sequences [51], the Euler-Maclaurin formula, or Voronoï's upper bound on the face counts of Voronoï cells of a lattice [121] are not well-known.

For the reader who wishes to refresh their familiarity with certain concepts, we provide some book references. For Chapter 2, we assume some rudimentary familiarity with Hamming space. The freely available book by O'Donnell [89] covers that and much more. For Chapter 3, we assume some familiarity with stastical signal processing, estimation, and elementary number theory. For the signal processing and estimation aspects, the book by Luenberger [80] covers all that we need and more at a level suitable for both mathematicians and engineers. There are countless sources for elementary number theory freely available online, such as the book by Stein [108]. For Chapter 4, we assume some familiarity with the basic notions of quantization. The excellent survey by Gray and Neuhoff [57] covers all that we need and more. It also has the virtue of tracing the history of the field accurately.

We apologize in advance to the mathematicians who desire a higher level of sophistication, and to the engineers who just want to build things and move on. Good examples of the balance we are striving towards are anything written by Donald Knuth, in particular the outstanding discrete mathematics book with Graham and Patashnik [56]. Naturally, we do not reach that level either, so we must apologize yet again.

A few words on notation. We assume knowledge of standard asymptotic notation $o(), O(), \Theta(), \Omega(), \omega()$ with their usual meanings. We simply use the word *constant* to refer to what many authors call a *universal constant*. We shall call scenarios with exactly matching upper and lower bounds *sharp*, and the analagous situations with upper and lower bounds that remain within a constant multiplicative factor of each other *tight*. We assume that the reader is familiar with the "indicator/characteristic function" notation $\mathbb{1}(x = 0), \mathbb{1}(x \in \mathcal{A})$. We sometimes find it convenient to follow the analytic number theorists and write

$$e(z) \triangleq e^{2\pi i z}.$$

Conclusions and directions for future research are provided on a chapter by chapter basis, once again in line with the self-contained philosophy. As such, our final Chapter 5 contains only certain general remarks about this dissertation and where one can cultivate ideas further.

Chapter 2

Linear Programming Bounds and Anticoodes in Hamming Space

> That combinatorics and information sciences often come together is no surprise - they were born as twins (Leibniz “Ars Combinatoria” gives credit to Raimundus Lullus from Catalania, who wanted to create a formal language).
>
> *Rudolf Ahlswede*, 2006 Shannon Lecture

2.1 Introduction

In coding theory, the basic question is to maximize the size of a set subject to a minimum distance requirement. The analogous *dual* question is to maximize the size of a set subject to a maximum distance constraint. This maximization problem may be termed as an *anticoding* or more precisely an *isodiametric* problem. We note that this *duality* can be made precise in terms of the *anticoding bound* of Delsarte [35, Thm 3.9]. This *anticoding bound* is sharper than the *sphere-packing* bound when a optimal solution to the isodiametric problem is not given by a ball.

In Hamming space, a complete resolution of the isodiametric problem was given by [7, Diametric Theorem], building upon techniques as well as the complete resolution of the analogous question for Johnson space given in [6]. In the non-binary setting of Hamming space, the optimal anticode is not in general a ball, but rather a Cartesian product of a ball and a subcube. In particular, this implies that the *anticoding bound* is sharper than the classical *sphere-packing* bound in the non-binary setting, as noted in [4].

Our primary goal in this chapter is to address anticoding problems in a complementary manner to the diametric perspective of [6, 7]. Specifically, note that in *isodiametric* problems the goal is to maximize the cardinality subject to a constraint that may be viewed as an upper bound on a potential energy characterized by a hard pair potential function. Our aim here is to "flip" this perspective, and ask energy minimization questions subject to a cardinality constraint. Although the two questions are obviously related (see e.g. (2.1)), we find the phenomena sufficiently rich (e.g. Theorems 2, 3, 4) to warrant investigation in their own right. We note that the complementary investigation of anticodes via potential energy may be traced in [5, pg vii,239]. There, the authors describe the question of the average cost of a uniformly randomly chosen update within a subset of a metric space. The authors then specialize to a cost function that decomposes on a product space as a sum of cost functions on the individual coordinates.[1] In general, taking a hard cost constraint with cost 0 for distances below a threshold, and ∞ otherwise leads one naturally to diametric problems. Our work may be viewed in that framework as considering other cost functions, with particular emphasis on completely monotone ones which we define in 4. A more direct motivation is viewing our work as addressing the anticoding analog of the *ground state* version of the coding problem as described in [29]. Along the way, we draw a connection with the problem of maximum noise stability in theoretical computer science, and thereby answer a folklore question of Mossel that we heard from Razenshteyn and Ramnarayan in Corollary 1. We use this perspective to guide our work in Section 2.5 onwards.

We first define what we mean by energy below.

Definition 1. *Let $(\mathcal{X}, d)$ be a finite metric space, and $f : \mathbb{R} \to \mathbb{R} \cup \{\pm\infty\}$ a* potential

[1] This perspective and a motivation for such study in terms of updating memories with cost constraints was also emphasized in Ahlswede's 2006 Shannon lecture [3].

function. Let $\mathcal{C} \subseteq \mathcal{X}$. We then define the potential energy *of $\mathcal{C}$ with respect to the* potential function f *to be*

$$E_f(\mathcal{C}) \triangleq \frac{1}{|\mathcal{C}|} \sum_{\substack{x,y \in C \\ x \neq y}} f(d(x,y)).$$

Then we may define two fundamental limits associated with the problem of finding energy minimizing (ground) states.

Definition 2.

$$e^*(f,c) \triangleq \min_{\mathcal{C} \subseteq \mathcal{X}: |\mathcal{C}|=c} E_f(\mathcal{C}).$$

$$c^*(f,e) \triangleq \max_{\mathcal{C} \subseteq \mathcal{X}: E_f(\mathcal{C}) \leq e} |\mathcal{C}|.$$

In general, sharp information about one of these functions does not necessarily translate into sharp information about the other function, even when one confines oneself to "interesting" potential functions, such as exponential decays. What is trivially clear however is that c^*, e^* are related by

$$c^*(f, e^*(f,c)) \geq c \tag{2.1a}$$

$$e^*(f, c^*(f,e)) \leq e. \tag{2.1b}$$

A natural question then is what constitutes an interesting potential function. One class of examples is readily furnished by the isodiametric problem, that is finding $\max_{\mathcal{C} \subseteq \mathcal{X}: d(\mathcal{C}) \leq d} |\mathcal{C}|$, where:

Definition 3. *The* diameter *of a set $\mathcal{C}$ in a finite metric space $(\mathcal{X}, d)$ is given by: $d(\mathcal{C}) \triangleq \max_{x,y \in \mathcal{C}} d(x,y)$.*

It is then clear that the isodiametric problem is nothing but the question of determining $c^*(f, |\mathcal{X}|)$ for

$$f(x) = \begin{cases} 1 & x \leq d \\ \infty & x > d. \end{cases}$$

Another class of functions, namely that of *completely monotonic* functions has proved to be very fruitful from a theoretical perspective in these investigations. Moreover, in the Euclidean setting, special cases of completely monotonic functions such as power laws have a natural physical interpretation. In coding theory, optimizing these potential functions gives information on the probability of error via the union bound.

In the discrete setting, completely monotonic functions are defined as follows:

Definition 4. *Let Δ denote the finite difference operator, defined by $\Delta f(n) \triangleq f(n+1) - f(n)$. Then, a function $f : \{a, a+1, \ldots, b\}$ is said to be* completely monotonic *if its iterated differences alternate in sign, that is $(-1)^k \Delta^k f(i) \geq 0$ whenever $k \geq 0$ and $a \leq i \leq b-k$.*

Of crucial importance for us is the fact that $f(r) = \gamma^r$ for $0 \leq \gamma \leq 1$ is a completely monotonic function. Minimizing potential energy with respect to such f favors repulsion between points, and is a ground state analog of the coding problem. A natural, perhaps naive view of anticoding is simply to flip the sign of f, or equivalently maximize the potential energy associated to such f.

What we find surprising is that this approach still retains an "operational significance" in the anticoding setting via a connection with the problem of *maximal noise stability* in theoretical computer science. Noise stability was first studied explicitly in [18]; see for instance [89, 2.4] for an introduction to the topic. Typically, noise stability is studied for Boolean functions $f : \{0,1\}^n \to \{0,1\}$. However, as the methods we employ apply more generally, we shall define an analogous notion for $f : \mathbb{F}_q^n \to \{0,1\}$. Although this notation suggests that $\mathbb{F}_q$ is a finite field, we shall not use the field structure.

As the name may suggest, noise stability is measured by $\Pr(f(x) = f(y))$ where y is a noisy version of x. Typically, one is interested in the behavior of functions on product spaces, and thus it is natural to consider product transition probability kernels, though it can be defined in greater generality.

We define it rigorously as follows in the context of Hamming space, and also define the problem of *maximal noise stability* subject to an expected value constraint.

Definition 5. *Let $f : \mathbb{F}_q^n \to \{0,1\}$ be a Boolean valued function. Let $r(\cdot|\cdot)$ be a row-stochastic $q \times q$ transition probability matrix. We define a kernel $s(\cdot|\cdot)$ on the product space $\mathbb{F}_q^n$. Typically, s is a product kernel given by $s(b^n|a^n) = \prod_{i=1}^n r(b_i|a_i)\ \forall a^n \in \mathbb{F}_q^n$. Let $\mathbf{x} \sim U(\mathbb{F}_q^n)$ be a uniformly distributed random variable. Let $\mathbf{y}$ be coupled with $\mathbf{x}$ by sending $\mathbf{x}$ through the kernel s. Then the* noise stability *of f is given by*

$$\mathbf{Stab}_s(n, f) \triangleq \Pr(f(\mathbf{x}) = f(\mathbf{y})).$$

We also define the maximal noise stability function *as*

$$\mathbf{Stab}_s^*(n, \mu) \triangleq \max_{f : \mathbb{E}[f(\mathbf{x})] = \mu} \mathbf{Stab}_s(n, f).$$

Since it is usually clear that we are referring to a product kernel, we will often simply write $\mathbf{Stab}_r(n, \mu)$. Furthermore, when r is parametrized in a natural way, such as a binary symmetric channel (BSC) with parameter ϵ, we may write $\mathbf{Stab}_\epsilon(n, \mu)$. Similar remarks apply to $\mathbf{Stab}_r^*, \mathbf{Stab}_\epsilon^*$.

We also find it convenient to define the notation

$$\mathbf{Stab}_s(n, \mathcal{C}) \triangleq \mathbf{Stab}_s(n, \mathbb{1}(x \in \mathcal{C})).$$

Note that we do not strictly follow the conventions of [89, pg 53], which defines $\mathbf{Stab}_s(n, f) \triangleq \mathbb{E}[f(\mathbf{x})f(\mathbf{y})]$. The reason for this is that the above definition is better suited for generality. Note also that it is common in the study of Boolean functions to work with $f : \{-1,1\}^n \to \{-1,1\}$; this simply corresponds to a relabeling $0 \leftrightarrow -1, 1 \leftrightarrow 1$. With the $\{-1,1\}$ output convention, it is clear that the definition of noise stability adopted in [89, pg 53] is simply an affine transformation of Definition 5. For the sake of notational clarity, in all rigorous statements we shall make it clear which representation we are working with.

2.2 Main results

We now give precise statements of the main results that we establish in this chapter. Lemmas, proofs, and establishment of statements of possible independent interest will occupy subsequent sections.

2.2.1 Linear programming and isodiametry in Hamming space

We first revisit in Section 2.3 the diametric theorem of [7] from a linear programming (LP) bound perspective, and rederive a sharp bound for the subcube cases. Our demonstration of the sharp cases of the LP bound for isodiametry in Hamming space may be viewed as an analog of the work of Wilson [124], who used LP to establish special cases of the complete diametric theorem obtained in [6]. We note that Shinkar [104] has also established the subcube cases using spectral techniques. Indeed, the spectral techniques used in [104] are implicitly contained in the language of association schemes and LP bounds of [35].

Definition 6. *A* subcube *of cardinality q^k in Hamming space $\mathbb{F}_q^n$ is defined by $\mathcal{C}_{q,k} = \{x : x_i = 0 \quad \forall 1 \leq i \leq n-k\}$.*

Theorem 1. *Let $N_q(n,d) \triangleq \max_{\mathcal{C} \subseteq \mathbb{F}_q^n : d(\mathcal{C}) \leq d} |\mathcal{C}|$. Then if $d \leq 1$ or $d \geq n-q+1$, $N_q(n,d) = q^d$, and this may be deduced from the LP bounds. Moreover, equality is attained by subcubes: $\mathcal{C} = \mathcal{C}_{q,d}$.*

2.2.2 Universal optimality for special subcubes

We next establish in Section 2.4 the fact that some special subcubes are simultaneous ground states in the anticoding sense for classes of potential functions. This phenomenon is called *universal optimality* as defined in [26]. However, it turns out that for some subcubes we get even stronger information than being universally optimal with respect to all completely monotonic functions, and in fact we can deduce universal optimality with respect to all monotonic functions. Most of Theorem 2 follows in a natural manner from the LP bounds, except for (2.6) that relies on a certain inequality for Krawtchouk polynomials established in Lemma 13.

Theorem 2. *Consider the class $\mathcal{F}$ of all nonnegative monotonically nondecreasing potential functions $f : \{0, 1, \ldots, n\} \to \mathbb{R}$, that is with $f(i) \leq f(i+1)$ $\forall 0 \leq i \leq n-1$, and $f(0) \geq 0$. Then $\forall f \in \mathcal{F}, q \geq 2 \in \mathbb{N}, n \geq 2$*

$$e^*(f, q) = E_f(\mathcal{C}_{q,1}) \tag{2.2}$$

$$e^*(f, q^2) = E_f(\mathcal{C}_{q,2}) \tag{2.3}$$

$$e^*(f, q^{n-1}) = E_f(\mathcal{C}_{q,n-1}) \tag{2.4}$$

$$e^*(f, q^n) = E_f(\mathcal{C}_{q,n}). \tag{2.5}$$

Furthermore, if $q > 2$, we have for all $f \in \mathcal{F}$ and $n \geq 2$

$$e^*(f, q^{n-2}) = E_f(\mathcal{C}_{q,n-2}). \tag{2.6}$$

Now consider the class $\mathcal{G}$ of all negations of completely monotonic potential functions $f : \{0, 1, \ldots, n\} \to \mathbb{R}$. In other words, $f \in \mathcal{G}$ iff $(-1)^{k+1}\Delta^k f(i) \geq 0$ whenever $k \geq 0$ and $0 \leq i \leq n-k$. Then for $q = 2$, we have for all $f \in \mathcal{G}$ and $n \geq 2$

$$e^*(f, 2^{n-2}) = E_f(\mathcal{C}_{2,n-2}). \tag{2.7}$$

As rather simple corollaries of Theorem 2, we obtain the following implications for the problem of maximal noise stability subject to an expected value constraint. Our deduction of these statements for maximal noise stability is based on a connection between anticoding and maximum noise stability that we develop in Section 2.5.

In binary Hamming space, we have the following.

Corollary 1. *Let $q = 2$, and let us work with functions $f : \{-1, 1\}^n \to \{-1, 1\}$. Let the transition probability kernel w be given by the family of BSC(ϵ). In other words,*

$w(1|1) = w(-1|-1) = 1-\epsilon$ and $w(-1|1) = w(1|-1) = \epsilon$. Let

$$g(x) = \frac{x_1x_2 + x_1 + x_2 - 1}{2} = x_1 \wedge x_2$$

where $\wedge$ denotes logical "and". Then $\forall 0 \leq \epsilon \leq \frac{1}{2}$ and $\forall n$ we have

$$\mathbf{Stab}^*_\epsilon\left(n, \frac{-1}{2}\right) = \mathbf{Stab}_\epsilon(n, g).$$

Similarly, we have

$$\mathbf{Stab}^*_\epsilon\left(n, \frac{1}{2}\right) = \mathbf{Stab}_\epsilon(n, -g).$$

We also have the (well known)

$$\mathbf{Stab}^*_\epsilon(n, 0) = \mathbf{Stab}_\epsilon(n, h)$$

where

$$h(x) = x_1.$$

Corollary 1 answers a “folklore” question of Mossel that we first heard of from Razenshteyn and Ramnarayan.

In non-binary Hamming space, we have the following.

Corollary 2. *Let $q > 2$, and let us work with functions $f : \mathbb{F}_q^n \to \{0, 1\}$. Let the transition probability kernel w be given by the family of q-SC(ϵ). In other words, $w(y|x) = 1 - \epsilon$ if $y = x$, and $w(y|x) = \frac{\epsilon}{q-1}$ otherwise. Let $g_1(x) = x_1$ and $g_2(x) = \mathbb{1}(x_1 = 0)\,\mathbb{1}(x_2 = 0)$. Then $\forall 0 \leq \epsilon \leq 1 - \frac{1}{q}$ and $\forall n$ we have*

$$\mathbf{Stab}^*_\epsilon\left(n, \frac{1}{q}\right) = \mathbf{Stab}_\epsilon(n, g_1),$$
$$\mathbf{Stab}^*_\epsilon\left(n, \frac{q-1}{q}\right) = \mathbf{Stab}_\epsilon(n, \overline{g_1}),$$

We also have

$$\begin{aligned}\mathbf{Stab}^*_\epsilon\left(n, \frac{1}{q^2}\right) &= \mathbf{Stab}_\epsilon(n, g_2),\\ \mathbf{Stab}^*_\epsilon\left(n, \frac{q^2-1}{q^2}\right) &= \mathbf{Stab}_\epsilon(n, \overline{g_2}),\end{aligned}$$

Here, $\overline{f}$ *denotes the logical complement of* f*.*

We note that the results for measures $\frac{1}{4}, \frac{1}{2}$ are in some sense anticipated by the work of [50], who show (in our language) the optimality of subcubes of measure $\frac{1}{4}, \frac{1}{2}$ for the cost function $f(x) = x$.

2.2.3 A mean value theorem for noise stability

We note that Theorem 2 and the corresponding noise stability corollaries 1, 2 refer to a q-SC channel. The LP bounds (or their SDP generalizations) do not apply to general channels, and we are therefore unable to give sharp answers for such channels, even ones coming from a product noise. However, in Section 2.6, we prove a channel comparison, and show that the maximum noise stability for a general channel can be compared with the corresponding quantity for a q-SC with appropriate noise level ϵ. All of the statements in Section 2.6 follow from a statement that we call a *mean value theorem for noise stability*:

Theorem 3. *Let Aut denote the group of distance preserving automorphisms of Hamming space* $\mathbb{F}_q^n$*. Define a group action of Aut on Boolean valued functions* $f : \mathbb{F}_q^n \to \{0, 1\}$ *by* $(\sigma f)(x) = f(\sigma x)$ *where* $\sigma \in Aut$*. Let* $s(\cdot|\cdot)$ *denote a* $q^n \times q^n$ *probability kernel. Let* $t(\cdot|\cdot)$ *be a "symmetrized version" of* s*, given by*

$$t(y|x) = \frac{1}{|Aut|} \sum_{\sigma \in Aut} s(\sigma y|\sigma x). \tag{2.8}$$

Then we have

$$\frac{1}{|Aut|} \sum_{\sigma \in Aut} \mathbf{Stab}_s(n, \sigma f) = \mathbf{Stab}_t(n, f). \tag{2.9}$$

2.2.4 Large n and balls versus subcubes

From the above discussion, it is clear that both Hamming balls and subcubes have a role to play in Hamming space for anticoding problems. In Section 2.7 we combine Hamming balls and subcubes by a *stacking* construction to prove that universal optima (in the sense of noise stability across the q-SC(ϵ) family) form a sparse set:

Theorem 4. *Let $\mathcal{S}$ denote the set of cardinalities $0 \leq c \leq q^n$ where there exists a universally optimal anticode $\mathcal{C}$ with $|\mathcal{C}| = c$, where the universal optimality is in the sense of noise stability across the q-SC(ϵ) with $\epsilon \in [0, 1 - \frac{1}{q}]$. Then, $\frac{|\mathcal{S}|}{q^n} = o(1)$ as $n \to \infty$.*

Along the way, we establish a rigorous definition of $\mathbf{Stab}_r^*(\infty, \mu)$ for any product noise generated by the kernel $r(\cdot|\cdot)$ in Proposition 5. The rigorous definition is meant to capture a *large* n limit. The Lemmas 15, 17 that we use to establish 5 also play a role in our proof of 4.

2.3 Linear programming and isodiametry in Hamming space

We now prove Theorem 1. As noted in Section 2.2, the theorem itself is completely subsumed by the complete diametric theorem of [7], and the subcube cases have been derived independently of us by spectral techniques in [104]. As such, the purpose of this section is to introduce the LP bounds in Hamming space which cover the techniques used in [104] and more importantly play a key role in the remainder of this chapter.

2.3.1 LP bounds and Fourier analysis on Hamming space

First, we formulate the LP bounds. Suppose

$$\mathcal{C} \subseteq \mathbb{F}_q^n.$$

The Delsarte bounds are linear constraints on the *distance distribution*

$$A_i \triangleq \frac{1}{|\mathcal{C}|}|\{(x,y) \in \mathcal{C} \times \mathcal{C} : d(x,y) = i\}|. \tag{2.10}$$

As we are working in Hamming space, here $d(x,y)$ is the Hamming distance between x and y.

We now define the Krawtchouk polynomials.

Definition 7.

$$\begin{aligned} K_k(x) = K_k(x;n) &= K_k(x;n,q) \\ &= \sum_{j=0}^{k} (-1)^j (q-1)^{k-j} \binom{x}{j}\binom{n-x}{k-j}. \end{aligned} \tag{2.11}$$

Then the Delsarte inequalities [35, Thm 3.3,4.2] are

$$\sum_{i=0}^{n} A_i K_j(i) \geq 0 \quad \forall 0 \leq j \leq n. \tag{2.12}$$

These inequalities are central in coding theory, and so we believe it is worth having a look at a proof of these inequalities and how they are connected to Fourier analysis. All of this exposition on the Delsarte inequalities is in some sense "classical" and is either explicitly or implicitly contained in his seminal work [35]. For readers who want a quick derivation of the inequalities themselves, we recommend [118, Sec. 5.3].

Perhaps it is useful to first understand where the Krawtchouk polynomials come from, as their definition 7 is relatively unilluminating. We have the following Lemma (see e.g. [118, Lemma 5.3.1]):

Lemma 1. *Let $\langle x, y\rangle$ denote the usual inner product in $(\mathbb{Z}/q\mathbb{Z})^n$. Let $\omega = e\left(\frac{1}{q}\right)$ be a primitive q^{th} root of unity. Let $x \in \mathbb{F}_q^n$ be a fixed word of weight i, in other words $|x| = i$. Then,*

$$\sum_{y \in \mathbb{F}_q^n, |y|=k} \omega^{\langle x,y\rangle} = K_k(i).$$

Proof of Lemma 1. By the underlying symmetries of Hamming space, we may assume without loss that $x = (x_1, x_2, \ldots, x_i, 0, 0, \ldots, 0)$, where $x_i \neq 0$. Choose k positions $h_1, h_2, \ldots, h_k$ with $0 < h_1 < h_2 < \cdots < h_j \leq i < h_{j+1} < \cdots < h_k \leq n$. Let $\mathcal{D}$ be the set of all words of weight k that have their nonzero coordinates in precisely these positions. Then, we have

$$\begin{aligned}
\sum_{y \in \mathcal{D}} \omega^{\langle x,y \rangle} &= \sum_{y_{h_1} \in \mathbb{F}_q \setminus \{0\}} \cdots \sum_{y_{h_k} \in \mathbb{F}_q \setminus \{0\}} \prod_{i=1}^{k} \omega^{x_{h_i} y_{h_i}} \\
&= \left[\prod_{i=1}^{j} \sum_{y \in F_q \setminus \{0\}} \omega^{x_{h_i} y} \right] (q-1)^{k-j} \\
&= \left[\prod_{i=1}^{j} \left(\omega + \omega^2 + \ldots \omega^{q-1} \right) \right] (q-1)^{k-j} \\
&= (-1)^j (q-1)^{k-j}.
\end{aligned}$$

Since there are $\binom{i}{j}\binom{n-i}{k-j}$ choices for $\mathcal{D}$, we get the result once we recall Definition 7.

□

The key role played by the roots of unity should already suggest the Fourier analytic nature of the linear programming bounds. Let us now develop Fourier analysis on the abelian group $(\mathbb{Z}/q\mathbb{Z})^n$. Once again, all this material is classical. For a reader who does not want to delve deeper into algebraic aspects such as representation theory and is more comfortable with analysis, we recommend [107, Ch. 7].

As remarked in Chapter 1, in the abelian case, it suffices to study certain scalar functions called *characters*:

Definition 8. *Let $(G, \cdot)$ be a finite abelian group, and let $S^1 = \{z \in \mathbb{C} : |z| = 1\}$ be the unit circle in the complex plane. A* character χ *on G is a complex valued function $\chi : G \to S^1$ which satisfies*

$$\chi(a \cdot b) = \chi(a)\chi(b).$$

In other words, it is a group homomorphism from G to S^1. The trivial character *(denoted by e) is given by $\forall g \in G, e(g) = 1$. Our notation here does collide with our notation $e(z) = e^{2\pi i z}$, but in practice the usage is unambiguous.*

The astute reader familiar with the DFT but not with characters should already realize that the DFT basis consisting of complex exponentials is a set of characters on $\mathbb{Z}/n\mathbb{Z}$.

A very important property of characters is that distinct characters are orthogonal. In order to show this, first we show

Lemma 2. *If χ is a nontrivial character for $(G, \cdot)$,*

$$\sum_{g \in G} \chi(g) = 0.$$

Proof of Lemma 2. As χ is nontrivial, there exists a b with $\chi(b) \neq 1$. Then,

$$\begin{aligned}\chi(b) \sum_{g \in G} \chi(g) &= \sum_{g \in G} \chi(b)\chi(g) \\ &= \sum_{g \in G} \chi(b \cdot g) \\ &= \sum_{g \in G} \chi(g),\end{aligned}$$

where the last step follows as $b \cdot g$ sweeps over all elements of G exactly once. Now as $\chi(b) \neq 1$, we must have the required $\sum_{g \in G} \chi(g) = 0$. □

We may now prove the important orthonormality of characters:

Proposition 1. *Let χ, χ' be two characters. Then $\langle \chi, \chi' \rangle = \mathbb{1}(\chi = \chi')$. Here*

$$\langle a, b \rangle \triangleq \frac{1}{|G|} \sum_{g \in G} a(g)\overline{b(g)}.$$

Proof of Proposition 1. First, if $\chi = \chi'$, then we have

$$\langle \chi, \chi' \rangle = \frac{1}{|G|} \sum_{g \in G} 1 = 1,$$

as required.

If $\chi \neq \chi'$, we have

$$\begin{aligned}
\langle \chi, \chi' \rangle &= \frac{1}{|G|} \sum_{g \in G} \chi(g) \overline{\chi'(g)} \\
&= \frac{1}{|G|} \sum_{g \in G} (\chi \chi'^{-1})(g) \\
&= 0.
\end{aligned}$$

Here, we used the fact that the characters themselves form a group (called the *dual group* $(\widehat{G}, \cdot)$) as is easily verified, and that $\chi\chi'^{-1}$ is a nontrivial character to which we can apply Lemma 2. □

Already we have some neat consequences, such as $|\widehat{G}| \leq |G|$ due to orthogonality implying independence and the fact that the dimension of the space of functions over G is $|G|$. In fact, we shall now show:

Theorem 5. *For a finite abelian group $(G, \cdot)$, $|G| = |\widehat{G}|$. In other words, the characters of a finite abelian group form a basis for the vector space of functions over G.*

Before turning to the proof of this nontrivial theorem, we do note that we do not strictly speaking need this to develop Fourier analysis for something as "concrete" as $(\mathbb{Z}/q\mathbb{Z})^n$. Indeed, the astute and impatient reader may already think: characters on product groups are just products of characters, and we know characters for $\mathbb{Z}/q\mathbb{Z}$ from knowledge of the DFT. More generally, one can invoke the structure theorem for finite abelian groups which allows one to decompose any such group into a direct product of cyclic groups, and in fact prove the above Theorem 5 from such a consideration.

However, we believe that with a little more patience more light may be shed. For example, going back to Chapter 1, we made some remarks about "simultaneous eigenfunctions for the translation group". The approach we present now elucidates this, assuming a basic grasp of linear algebra.

We first prove that commuting unitary operators on a finite dimensional inner product space are simultaneously diagonalizable:

Lemma 3. *Suppose $\{T_1, \ldots, T_k\}$ is a commuting family of unitary operators on the finite dimensional inner product space V; that is for all i, j*

$$T_i T_j = T_j T_i.$$

Then T_i are simultaneously diagonalizable. In other words, there is a basis for V consisting of eigenvectors for every $T_i, 1 \leq i \leq k$.

Proof. We shall induct on k. The case $k = 1$ is simply the spectral theorem. Suppose Lemma 3 is true for $k - 1$ commuting unitary operators. Applying the spectral theorem to T_k, we see that

$$V = V_{\lambda_1} \oplus \ldots V_{\lambda_s},$$

where V_{λ_i} denotes the subspace of all eigenvectors with eigenvalue λ_i. We now claim that each one of $T_1, \ldots, T_{k-1}$ maps V_{λ_i} to itself. For if $v \in V_{\lambda_i}$, and $1 \leq j \leq k - 1$, then we have

$$T_k T_j(v) = T_j T_k(v) = T_j(\lambda_i v) = \lambda_i T_j(v),$$

and so $T_j(v) \in V_{\lambda_i}$ as needed.

Now, the restrictions of $T_1, T_2, \ldots, T_{k-1}$ to V_{λ_i} are clearly well defined operators, and inherit pairwise commuting and the unitary nature. Furthermore, by the induction hypothesis, these are simultaneously diagonalizable. Thus, we get a suitable basis for each V_{λ_i} that works for $T_1, T_2, \ldots, T_k$. As V is a direct sum of the V_{λ_i}, we are done. □

With the above Lemma 3, we may now turn to the proof of Theorem 5.

Proof of Theorem 5. Define "translation" operators on the vector space of complex valued functions on G

$$(T_a f)(x) \triangleq f(a \cdot x).$$

G's abelian nature implies that T_a, T_b commute. Furthermore, T_a is clearly unitary, since $a \cdot x$ sweeps across G when x does. Then by Lemma 3 T_a is simultaneously diagonalizable. Thus, we have a basis of eigenvectors $v_b(\cdot)$, where b varies over G.

Let v be one of these v_b, and we claim that $w(g) \triangleq \frac{v(g)}{v(1)}$ is well-defined and also a character.

First, we need to show that $v(1) \neq 0$. Suppose not. Then

$$v(a) = v(a \cdot 1) = (T_a v)(1) = \lambda_a v(1) = 0,$$

forcing $v(a) = 0$ for an arbitrary a, which is impossible.

Thus w is indeed well-defined. We now check that it is indeed a character by

$$w(a \cdot b) = \frac{v(a \cdot b)}{v(1)} = \frac{\lambda_a v(b)}{v(1)} = \lambda_a \lambda_b = w(a)w(b).$$

This completes the proof of Theorem 5. □

Note that the above proof is quite hands-on and “effective”, since it gives an actual procedure (compute eigenvectors of the translation operators) to obtaining the characters. As such, the following consequence should be transparent either by the above discussion, or by the earlier remarks regarding characters on product groups being products of characters:

Corollary 3. *Let $\omega = e\left(\frac{1}{q}\right)$. The characters of $((\mathbb{Z}/q\mathbb{Z})^n, +)$ are χ_x, where $\chi_x(z) = \omega^{\langle x,z\rangle}$.*

Proof of Corollary 3. It is obvious that $\chi_x(z)$ are characters, since their range is on the unit circle, and

$$\chi_x(z_1 + z_2) = \omega^{\langle x,z_1+z_2\rangle} = \omega^{\langle x,z_1\rangle}\omega^{\langle x,z_2\rangle} = \chi_x(z_1)\chi_x(z_2).$$

Furthermore, we have one character for each of the elements of $(\mathbb{Z}/q\mathbb{Z})^n$, so we are done by Theorem 5. □

We now introduce the notion of a *positive definite function.*

Definition 9. *Let $(G, +)$ be a finite abelian group, and $f : G \to \mathbb{C}$. f is called positive definite iff for all N and $x_1, \ldots, x_N \in G$, $c_1, \ldots, c_N \in \mathbb{C}$,*

$$\sum_{j,k=1}^{N} f(x_j - x_k) c_j \overline{c_k} \geq 0.$$

In other words, for all N and all "codes" $x_1, \ldots, x_N$, the matrix A given by $A_{j,k} = f(x_j - x_k)$ is a positive definite matrix.

Before developing the theory further, let us look at some simple illustrations and consequences of Definition 9. For example, the function $f(x) = \mathbb{1}(x = 0)$ is a positive definite function; this simply follows from the fact that the identity is positive definite.

We also have the following simple

Lemma 4. *Let f be a positive definite function. Then*

1. *$f(0) \geq 0$,*

2. *$\forall x \in G, \quad f(0) \geq |f(x)|$,*

3. *$\forall x \in G, f(-x) = \overline{f(x)}$.*

Proof of Lemma 4. First take $N = 1, x_1 = 0, c_1 = 1$ to get $f(0) \geq 0$. Next take $N = 2, x_1 = 0, x_2 = x, c_1 = 1, c_2 = c$ where c is an arbitrary complex number on the unit circle, to get

$$\overline{c} f(-x) + c f(x) + 2 f(0) \geq 0.$$

Thus $f(-x) = \overline{f(x)}$ (the third item), and also $2\Re(cf(x)) \leq 2f(0)$. Take c proportional to $f(-x)$ to get the second item. □

It is also easy to show that characters are positive definite, and that positive definite functions form a cone:

Lemma 5. *Let $\widehat{g} \in \widehat{G}$ be a character. Then $\widehat{g}$ is positive definite. Also, a nonnegative linear combination of positive definite functions is positive definite. In other words, positive definite functions form a cone.*

Proof of Lemma 5. We have

$$\sum_{j,k=1}^{N} \widehat{g}(x_j - x_k)c_j\overline{c_k} = \sum_{j,k=1}^{N} \widehat{g}(x_j)\overline{\widehat{g}(x_k)}c_j\overline{c_k} = \left|\sum_{j=1}^{N} \widehat{g}(x_j)c_j\right|^2 \geq 0,$$

completing the proof of the first statement by the definition of positive definiteness 9. The second statement is obvious, and follows directly from linearity. □

We may now prove a (weak) form of Bochner's theorem (see e.g. [22] for original, or an exposition [70, VI.2.7] for the version on the locally compact $\mathbb{R}/\mathbb{Z}$.)[2] The "weakness" here refers to the fact that we are working in the technically simple setting of finite abelian groups.

Theorem 6 (Bochner)**.** *Let $(G,+)$ be a finite abelian group, and let $f : G \to \mathbb{C}$ have Fourier expansion*

$$f(x) = \sum_{\widehat{g}\in\widehat{G}} a_{\widehat{g}}\widehat{g}(x).$$

Then, f is positive definite iff all its Fourier coefficients $a_{\widehat{g}}$ are nonnegative.

Proof of Theorem 6. Let $\widehat{h} \in \widehat{G}$ be a character. Now take $x_1, x_2, \ldots, x_N = G$ (each element of G occuring precisely once), and $c_j = \overline{\widehat{h}(x_j)}$. Applying the definition of positive definiteness 9, we have

$$\begin{aligned}
&\sum_{\widehat{g}\in\widehat{G}} a_{\widehat{g}} \sum_{j,k} \widehat{g}(x_j - x_k)\overline{\widehat{h}(x_j)}\widehat{h}(x_k) \geq 0\\
&\Rightarrow \sum_{\widehat{g}\in\widehat{G}} a_{\widehat{g}} \sum_{j,k} \widehat{g}(x_j)\overline{\widehat{g}(x_k)}\overline{\widehat{h}(x_j)}\widehat{h}(x_k) \geq 0\\
&\Rightarrow \sum_{\widehat{g}\in\widehat{G}} a_{\widehat{g}} \left|\sum_{j} \widehat{g}(x_j)\overline{\widehat{h}(x_j)}\right|^2 \geq 0\\
&\Rightarrow |\widehat{G}|a_{\widehat{h}} \geq 0\\
&\Rightarrow a_{\widehat{h}} \geq 0,
\end{aligned}$$

[2] It can be argued that these ideas for $\mathbb{R}/\mathbb{Z}$ should be traced further back to the work of Herglotz [62] and independently Riesz [96].

where we used the orthonormality of characters (Proposition 1). $\widehat{h}$ was arbitrary, so we proved one direction.

We now need to show that if all the Fourier coefficients are nonnegative, f is positive definite. But this is precisely the content of Lemma 5. □

Most of the above machinery is not strictly speaking needed for the proof of the central Delsarte inequalities (2.12). However, it does serve to place them in a broader Fourier analytic context, and helps reduce the "mystery" of where the inequalities come from. Let us now prove the Delsarte inequalities.

Proposition 2. *Let the Krawtchouk polynomials be defined via* (2.11)*, and let the distance distribution of a code $\mathcal{C}$ be given by A_i, defined by* (2.10)*. Then, we have* (2.12)

$$\forall 0 \leq j \leq n, \quad \sum_{i=0}^{n} A_i K_j(i) \geq 0.$$

Proof of Proposition 2. Let $x_1, \ldots, x_N = \mathcal{C}$, where each x_i occurs precisely once for each element of the code $\mathcal{C}$. Let $c_i = 1$ for all i. χ_x is positive definite, so $\sum_{x:|x|=j} \chi_x$ is positive definite as well by Lemma 5. The definition of positive definiteness 9, together with the expression of Krawtchouk polynomials in terms of characters (Lemma 1) completes the proof. More explicitly, we have

$$\begin{aligned}|\mathcal{C}| \sum_{i=0}^{n} A_i K_j(i) &= \sum_{i=0}^{n} \sum_{(x,y) \in \mathcal{C}^2, |x-y|=i} \left[\sum_{z \in \mathbb{F}_q^n, |z|=k} \omega^{\langle x-y,z \rangle} \right] \\ &= \sum_{z \in \mathbb{F}_q^n, |z|=k} \left| \sum_{x \in \mathcal{C}} \omega^{\langle x,z \rangle} \right|^2 \geq 0.\end{aligned}$$

□

Remark 1. *The above proof also reveals an interpretation of the Delsarte inequalities. Consider $f(x) = \mathbb{1}(x \in \mathcal{C})$, that is the "characteristic function" of the code. Then, the Delsarte inequalities state that the sum of the squares of the magnitudes of the Fourier coefficients of f, grouped by Hamming weight j, are nonnegative. For the* *reader familiar with Boolean Fourier analysis in the spirit of [89], this provides a link to the Delsarte inequalities.*

Remark 2. *The astute reader may have noticed that nothing in our above discussion really sheds light as to why the Krawtchouk polynomials are actually polynomials, though we did verify this by direct computation. As our direct computation in the proof of Lemma 1 suggests, one may suspect that the polynomial behavior arises somehow from the combinatorics (in particular the regularity) of the underlying Hamming space. This guess is indeed correct, and there is a rich theory of association schemes that studies this further. Delsarte's original work [35] is written in that framework. We also refer the interested reader to the expository paper [36]. Although we could provide an exposition here as well, the theory of association schemes is very tangential to our focus here, and we feel that it would not shed much light on the subject matter at hand.*

We shall need a few properties of Krawtchouk polynomials for the subsequent development in Hamming space. The above remark illustrates how one may develop the theory systematically. Here, we shall content ourselves with a more "ad-hoc" development.

First, we obtain a closed form for the generating function of Krawtchouk polynomials (see e.g [118, 1.2.3]).

Lemma 6. *Recall (7)*

$$K_k(x) = \sum_{j=0}^{k} (-1)^j (q-1)^{k-j} \binom{x}{j} \binom{n-x}{k-j}.$$

Then we have

$$\sum_{i=0}^{\infty} K_i(x) z^i = (1+(q-1)z)^{n-x}(1-z)^x. \tag{2.13}$$

Proof of Lemma 6. Compare the coefficient of z^k on both sides of (2.13). The left hand side has $K_k(x)$, while on the right hand side we may use the binomial theorem to get a coefficient of

$$(-1)^j (q-1)^{k-j} \binom{x}{j} \binom{n-x}{k-j},$$

as required. □

Next, we obtain an orthogonality relation for Krawtchouk polynomials (see e.g. [36, Thm 1]).

Lemma 7. *We have*

$$\sum_{i=1}^{n} K_i(j) = q^n \, \mathbb{1}(j=0) - 1. \tag{2.14}$$

Proof of Lemma 7. It is obvious by the definition of Krawtchouk polynomials 7 that $K_0(j) = 1$, so we may rewrite the desired (2.14) as

$$\sum_{i=0}^{n} K_i(j) = q^n \, \mathbb{1}(j=0).$$

Let x be a fixed vector of Hamming weight j. We may use the expression of Krawtchouk polynomials in terms of characters (Lemma 1) to obtain

$$\begin{aligned}\sum_{i=0}^{n} K_i(j) &= \sum_{i=0}^{n} \sum_{y \in \mathbb{F}_q^n, |y|=i} \omega^{\langle x,y \rangle} \\ &= \sum_{y \in \mathbb{F}_q^n} \omega^{\langle x,y \rangle} \\ &= q^n \, \mathbb{1}(j=0),\end{aligned}$$

where for the last line we used a basic property of characters (Lemma 2). □

2.3.2 Proof of Theorem 1

With these inequalities in hand, we may turn to the proof of Theorem 1.

Proof of Theorem 1. We may first dispose of trivialities $d = n$ (the full space), and $d = 0$ (a single point). By (2.12), we see that $N_q(n, d) - 1$ is less than or equal to the solution of the following LP

$$\begin{aligned} \max_{A_i} \quad & \sum_{i=1}^{n} A_i \qquad (2.15) \\ \text{s.t.} \quad & A_i = 0 \quad \forall d < i \leq n \\ & A_i \geq 0 \quad \forall 1 \leq i \leq n \\ \sum_{i=1}^{n} & A_i K_j(i) \geq -K_j(0) \quad \forall 1 \leq j \leq n. \end{aligned}$$

The subtraction of 1 is simply due to $A_0 = 1$, and hence can be removed from the LP.

By weak duality, the primal LP given by (2.15) is upper bounded by the solution of the dual LP

$$\begin{aligned} \min_{p_i} \quad & \sum_{i=1}^{n} K_i(0) p_i \qquad (2.16) \\ \text{s.t.} \quad & p_i \geq 0 \quad \forall 1 \leq i \leq n \\ \sum_{i=1}^{n} & p_i K_i(j) \leq -1 \quad \forall 1 \leq j \leq d. \end{aligned}$$

The goal now is to construct dual variables p_i that yield the optimum cost $q^d - 1$ under the assumed constraints on d, and check their satisfiability. The way we will achieve this is by making the p_i satisfy a linear recurrence of order $n - d + 1$ with initial conditions

$$\begin{aligned} p_i &= 0 \quad \forall 1 \leq i \leq n - d \\ p_{n-d+1} &= \frac{1}{(q-1)^{n-d}}. \end{aligned}$$

Note that for $d = 1$, this completely specifies the dual variables; with no need to define the recurrence relation. For the $n - q + 1 \leq d \leq n - 1$ case, one considers a recurrence relation with characteristic polynomial $(z + \frac{1}{q-1})^{n-d}(z - 1)$. Explicitly, this yields the recurrence

$$p_k = \sum_{i=1}^{n-d+1} a_i p_{k-i}$$

where the weights a_i are

$$a_i = \frac{\binom{n-d}{i-1}}{(q-1)^{i-1}} - \frac{\binom{n-d}{i}}{(q-1)^i} \tag{2.17}$$

and thus satisfy $\sum_{i=1}^{n-d+1} a_i = 1$ by telescoping.

More importantly for our proof is the fact that $a_i \geq 0 \ \forall i$, which follows from (2.17), crucially using the assumption that $n-q+1 \leq d$. This yields immediately by induction that $p_i \geq 0 \ \forall i$.

By the general theory of linear recurrences, we know that

$$p_i = a + \left(\sum_{j=0}^{n-d-1} b_j i^j\right)\left(\frac{1}{1-q}\right)^i$$

for some constants a, b_j (depending on n, d, q but independent of i). It will be of convenience to switch to the falling factorial basis (in order to use Newton interpolation), and reindex by defining $q_i = p_{i+1}$, yielding

$$q_i = c + \left(\sum_{j=0}^{n-d-1} d_j \binom{i}{j}\right)\left(\frac{1}{1-q}\right)^i$$

for some constants c, d_j. The constants c, d_j are determined by have the specified boundary conditions

$$q_0 = q_1 = \cdots = q_{n-d-1} = 0, q_{n-d} = \frac{1}{(q-1)^{n-d}}.$$

Transposing and using the fact that $\Delta^k[(1-q)^n](0) = (-q)^k$ along with $\Delta^k[\binom{n}{j}] =$

$\binom{n}{j-k}$, one gets

$$d_i = -c(-q)^i \quad \forall 1 \leq i \leq n-d-1.$$

Using $q_0 = 0$, we get $d_0 = -c$, and hence

$$q_i = c\left[1 - \left(\sum_{j=0}^{n-d-1} (-q)^j \binom{i}{j}\right) \frac{1}{(1-q)^i}\right].$$

From $q_{n-d} = \frac{1}{(q-1)^{n-d}}$ and the binomial theorem, we can finally determine $c = \frac{1}{q^{n-d}}$, yielding the explicit formula for the dual variables

$$p_i = \frac{1}{q^{n-d}}\left[1 - \left(\sum_{j=0}^{n-d-1} \binom{i-1}{j} (-q)^j\right) \frac{1}{(1-q)^{i-1}}\right] \tag{2.18}$$

Nonnegativity of the p_i has been verified above, so it remains to compute the dual objective value and verify feasibility. For this purpose, we prove the following

Lemma 8.

$$\sum_{i=1}^{n} \binom{i-1}{s} K_i(j) \frac{1}{(1-q)^i} = (-1)^{s+1} \quad \forall 0 \leq s \leq n-j-1. \tag{2.19}$$

Proof of Lemma 8. The proof is by induction on s.

For $s = 0$, we may write (2.19) as

$$\sum_{i=1}^{n} K_i(j) \frac{1}{(1-q)^i} = -1$$

or equivalently

$$\sum_{i=0}^{\infty} K_i(j) \frac{1}{(1-q)^i} = 0.$$

But this follows immediately from the generating function governing Krawtchouk polynomials (2.13). Now suppose (2.19) holds for $s \leq k$, and we wish to prove it for $s = k+1$ where $k+1 \leq n-j$. Using $\binom{i-1}{k+1} = \binom{i}{k+1} - \binom{i-1}{k}$, we reduce to showing that

$$\sum_{i=1}^{n} \binom{i}{k+1} K_i(j) \frac{1}{(1-q)^i} = 0$$

or equivalently

$$\sum_{i=0}^{\infty} \binom{i}{k+1} K_i(j) \frac{1}{(1-q)^i} = 0 \tag{2.20}$$

Observe that as $k+1 \leq n-j$, $\binom{i}{k+1}$ is a polynomial in i of degree at most $n-j$. Thus, taking $x = j$, differentiating (2.13) at most $n-j$ times, and using the fact that $\frac{1}{1-q}$ is a root of order $n-j$, we get (2.20) as desired. □

We now turn to checking the dual LP's objective value for p_i governed by (2.18). First, note that

$$\sum_{i=1}^{n} K_i(j) = \sum_{i=0}^{n} K_i(j) K_0(i) - 1 \tag{2.21}$$

$$= q^n \, \mathbb{1}(j=0) - 1 \tag{2.22}$$

by an orthogonality relation of Krawtchouk polynomials (Lemma 7). Using (2.21) along with Lemma 2.19, we get

$$\sum_{i=1}^{n} p_i K_i(0) = \frac{1}{q^{n-d}} \left[q^n - 1 + (1-q) \left(\sum_{j=0}^{n-d-1} q^j \right) \right]$$
$$= q^d - 1,$$

as desired.

Turning to the verification of dual feasibility, we have

$$\sum_{i=1}^{n} p_i K_i(j) = \frac{1}{q^{n-d}} \left[-1 + (1-q) \left(\sum_{j=0}^{n-d-1} q^j \right) \right]$$
$$= -1 \quad \forall 1 \leq j \leq d,$$

as required.

It is obvious that subcubes achieve this objective value, thus completing the proof.

□

Note that this is an example of a sharp bound on the value of $c^*(f, q^n)$ for specific choices of f by the discussion in Section 2.1. In general, it seems like the LP bounds tend to yield richer information for the problem of $e^*(f, c)$; at the very least we know of many more examples where the LP bounds are sharp for $e^*(f, c)$. Apart from Theorem 1 and the work of [124, 49], where sharp answers are given for specific cases of the Johnson graph and for all cases of their q-analog, namely the Grassman graph respectively, we do not know of any other interesting families of finite distance-regular graphs where LP bounds yield sharp information for the isodiametric problem, although we do think this is highly plausible.

The remainder of this chapter will focus on some $e^*(f, c)$ questions. Before turning to specific choices of f, c, we note that one can without loss restrict study to anticodes that are *down-set* (for any q) and *right-compressed* (when $q = 2$) as long as f is a monotone potential. The arguments here are classical, with *right compression* operations going back to Erdős-Ko-Rado [42] and *pushing down* arguments that are essentially due to Kleitman [71].

It is convenient to define the notions of *slice* and *projection* of a set for this purpose, as in [6]. We also define a natural lexicographic order on $\mathcal{X}_q^n$.

Definition 10. *Let $\mathcal{C} \subseteq \mathbb{F}_q^n$, and let $\mathcal{J} \subseteq [n]$. For $x \in \mathbb{F}_q^n$, denote by x^J the subsequence of x obtained by deleting components x_t for $t \notin \mathcal{J}$. We then denote the* slice

$$\mathcal{C}_{\mathcal{J}}(x^{[n]\setminus\mathcal{J}}) \triangleq \{x^{\mathcal{J}} \in \mathbb{F}_q^{\mathcal{J}} : x \in \mathcal{C}\} \quad \textit{for } x^{[n]\setminus\mathcal{J}} \in \mathbb{F}_q^{[n]\setminus\mathcal{J}}.$$

We then denote the projection

$$\mathcal{C}_{\mathcal{J}} \triangleq \bigcup_{x^{[n]\setminus\mathcal{J}} \in \mathbb{F}_q^{[n]\setminus\mathcal{J}}} \mathcal{C}_{\mathcal{J}}(x^{[n]\setminus\mathcal{J}}).$$

We also denote by $\mathbf{L}$ a natural lexicographic order generated by $0 \leq 1 \cdots \leq q-1$ on $\mathbb{F}_q^n$, with the least significant positions towards the right end. We denote the set of the lexicographically first m elements in $\mathbb{F}_q^{\mathcal{J}}$ by $\mathbf{L}(\mathbb{F}_q^{\mathcal{J}}, m)$, and call this a lex-set.

With these notations in hand, we may readily define the *pushing down* operations as follows, again following [6].

Definition 11. *The* pushing down *operation with respect to the order* $\mathbf{L}$*, a subset* $\mathcal{J}$ *of indices, and an anticode* $\mathcal{C}$ *is given by*

$$D_{\mathcal{J}}(\mathbf{L}, \mathcal{C}) \triangleq \bigcup_{x^{[n]\setminus\mathcal{J}} \in \mathcal{C}_{[n]\setminus\mathcal{J}}} \{y : y^{[n]\setminus\mathcal{J}} = x^{[n]\setminus\mathcal{J}} \text{ and } y^{\mathcal{J}} \in \mathbf{L}(\mathbb{F}_q^{\mathcal{J}}, |\mathcal{C}(x^{[n]\setminus\mathcal{J}})|)\}$$

Then $\mathcal{C}$ *is said to be* down-set *if* $D_{\{i\}}(\mathbf{L}, \mathcal{C}) = \mathcal{C}$ *for all* $1 \leq i \leq n$*.*

In the special case of $q = 2$, we also define the *right compression* operations as follows.

Definition 12. *For any* $\mathcal{C} \in \mathbb{F}_2^n$*, any* $c = (c_1, \ldots, c_n) \in \mathcal{C}$*, and* $1 \leq i < j \leq n$ *we define the* right compressing *operators* $S_{i,j}$ *by*

$$S_{i,j}(c) \triangleq \\ \begin{cases} (c_1, \ldots, c_{i-1}, 0, c_{i+1}, \ldots, c_{j-1}, 1, c_{j+1}, \ldots, c_n), \text{if not in } \mathcal{C} \text{ and } c_j = 0, c_i = 1 \\ c \text{ otherwise.} \end{cases}$$

Also, let

$$S_{i,j}(\mathcal{C}) \triangleq \{S_{i,j}(c) : c \in \mathcal{C}\}.$$

Then $\mathcal{C}$ *is said to be* right-compressed *if* $S_{i,j}(\mathcal{C}) = \mathcal{C}$ *for all* $1 \leq i < j \leq n$*.*

With these definitions 11, 12, we may prove the following

Proposition 3. *Let* f *be a monotonically nondecreasing potential function, that is* $f \in \mathcal{F}$ *or in other words* $f(i) \leq f(i+1)$ $\forall 0 \leq i \leq n-1$*. Let a cardinality* c *with* $0 \leq c \leq q^n$ *be given. Then there exists a* down-set $\mathcal{C} \subseteq \mathbb{F}_q^n$ *such that for any other set* $\mathcal{C}'$ *of the same cardinality, we have*

$$E_f(\mathcal{C}) \leq E_f(\mathcal{C}').$$

Furthermore, if $q = 2$, we may take $\mathcal{C}$ to be simultaneously down-set *and* right-compressed.

Proof of Proposition 3. Suppose $\mathcal{C}_0$ minimizes $E_f(\mathcal{A})$ over all anticodes $\mathcal{A}$ with $|\mathcal{A}| = c$. Consider the sets $\mathcal{C}_i = D_{\{i \pmod n)+1\}}(\mathbf{L}, \mathcal{C}_{i-1})$ defined inductively for $i \geq 1$. It is clear that this process stabilizes eventually, and moreover that it does not change the cardinality. For example, consider a bounded potential

$$F(\mathcal{C}) \triangleq \sum_{c \in \mathcal{C}} \overline{c}_q, \tag{2.23}$$

where $\overline{c}_q = \sum_{i=0}^{n-1} c_{n-i} q^i$. This potential clearly can not increase with i, so it eventually stabilizes. Once the potential stabilizes, it is clear that any further iterations do not change the set. Let us denote this stabilized *down-set* by $\mathcal{D}$. We argue now that $E_f(\mathcal{C}_0) \geq E_f(\mathcal{D})$. Clearly it suffices by symmetry to show that $E_f(D_{\{n\}}(\mathbf{L}, \mathcal{C})) \leq E_f(\mathcal{C})$ for any $\mathcal{C}$. But this is easy to see. Consider two arbitrary slices of the anticode

$$\mathcal{A} = x^{[n]\setminus\{n\}} \times \mathcal{C}_{\{n\}}(x^{[n]\setminus\{n\}}),$$
$$\mathcal{B} = y^{[n]\setminus\{n\}} \times \mathcal{C}_{\{n\}}(y^{[n]\setminus\{n\}}),$$

and note that the action of D on $\mathcal{A}, \mathcal{B}$ can not increase the number of pairs of codewords of $\mathcal{A}, \mathcal{B}$ of Hamming distance $\geq i$ for any $0 \leq i \leq n$. $E_f(\mathcal{C}_0) \geq E_f(\mathcal{D})$ then follows by the trivial spanning set characterization of $\mathcal{F}$ established in Lemma 10. Taking $\mathcal{C} = \mathcal{D}$ completes the proof for $q \neq 2$.

Now suppose $q = 2$. Let $\mathcal{D}_0 = \mathcal{D}$, $\mathcal{D}_i = S_{kl(i)}(\mathcal{D}_{i-1})$ defined inductively for $i \geq 1$. Here, $kl(i)$ denotes some periodically repeating indexing of all (k, l) pairs from $1 \leq k < l \leq n$. Once again, it is clear that this process eventually stabilizes and that it does not change the cardinality; indeed the same potential (2.23) works as a certificate. Let $\mathcal{C}$ denote the stabilized *right-compressed* set. It remains to check that $\mathcal{C}$ is also *down-set* and that the number of pairs of codewords with Hamming distance $\geq i$ does not increase for any $0 \leq i \leq n$. Again by symmetry it suffices to study

$S_{n-1,n}$. Consider two arbitrary slices of the anticode

$$\mathcal{A} = x^{[n]\setminus\{n,n-1\}} \times \mathcal{C}_{\{n,n-1\}}(x^{[n]\setminus\{n,n-1\}}),$$
$$\mathcal{B} = y^{[n]\setminus\{n,n-1\}} \times \mathcal{C}_{\{n,n-1\}}(y^{[n]\setminus\{n,n-1\}}),$$

and note by a case by case analysis of the bits in positions $\{n-1,n\}$ of $\mathcal{A}$ that $S_{n-1,n}$ preserves the *down-set* property. Moreover, by looking at both $\mathcal{A}, \mathcal{B}$, it is clear that the number of pairs of codewords with Hamming distance past a threshold does not increase. This completes the proof for $q = 2$. □

We remark that the powerful *pushing-pulling* operations that serve a crucial role in establishing the complete diametric theorems of [6, 7] in general modify the cardinality of the anticode, unlike the classical *compression* and *pushing down* operations. *Pushing-pulling* thus seem better suited to $c^*(f, e)$ questions or generalizations with a non-uniform ground measure such as a Bernoulli weighted case (see e.g. [46] and references therein). As our focus here is on $e^*(f, c)$ questions, we do not examine this further.

2.4 Universal optimality of some subcubes

Before we turn to the proof of Theorem 2, it is natural to wonder why we focus on subcubes. For example, we know by the diametric theorem [7] that Hamming balls or nontrivial Cartesian products of Hamming balls and subcubes are optimal anticodes in the diametric sense, with pure subcubes addressing only a small range of the diametric problem.

The reason for this is because of our focus on the phenomenon of universal optimality. Consider a potential function $f(x) = -\binom{n-x}{n-1}$. $-f$ is completely monotonic, and the problem of $e^*(f, c)$ is nothing but the *edge-isoperimetric* problem in Hamming space. In the binary case this was first solved by Harper [60] and in the non-binary case by Lindsey [76]. In either case, the answer is given by taking the first c vertices in lexicographic order in Hamming space. Furthermore, this answer is unique up to symmetry, specifically the group of (Hamming) distance preserving automorphisms.

Thus, for the problem of universal optimality, where the class of functions includes one that minimizes the edge boundary (such as completely monotonic functions or monotonic functions), we can focus on such lex-sets, which we can denote formally using our notation as $\mathbf{L}(\mathbb{F}_q^n, m)$.

Of course, this does not address why we focus on the further specialization to subcubes. Our restricted focus on subcubes comes from a limitation of the LP bounds observed in [29, Prop. 29]. Suppose our set of potential functions contains $n+1$ linearly independent potential functions; this is true of the cones of completely monotonic and monotonic functions. Then by [29, Prop. 29] we know that $\mathcal{C} \setminus \{c\}$ must have the same distance distribution for all $c \in \mathcal{C}$. If $\mathcal{C}$ is a lex-set, this can only happen when $\mathcal{C}$ is a subcube. To see this, consider $c_1 = 0, c_2 = m$ where m is the maximum element in $\mathcal{C}$ in lexicographic order, and look at how many neighbors each of them has at Hamming distance 1. The only way these equate is if c_2 is the maximum element of a subcube in lexicographic order. It would be interesting if one could get around this limitation (for example, using semidefinite programming (SDP) bounds of some constant order [102, 53]) and prove an analog of (2.7) for cardinality $(3/8)2^n$, or other candidate cardinalities outlined later in this chapter, such as the presumably easier Conjecture 1 which asks the question for $(1/8)2^n, (1/16)2^n$.

We now turn to the proof of Theorem 2. With a good exact LP solver that supports rational arithmetic (we used [54], which incorporates exact rational LP support from [55]) and a reasonable ability to identify sequences, for instance by looking at finite differences, it is possible to "eye-ball" families of dual certificates of optimality and prove Theorem 2. However, by stepping back and taking a more conceptual view, one can get a better feel for what is happening. Accordingly, the proof we give here uses slightly more machinery about the LP bound, most of which can be found in [29].

Following [29], we use the term *quasicode* for a feasible point in the Delsarte LP [29, Defn 6].

Definition 13. *A* quasicode $\mathbf{a}$ *of length n and size N over $\mathbb{F}_q$ is a real column vector*

$(A_0, A_1, \ldots, A_n)$ *satisfying the Delsarte inequalities* (2.12)*. Explicitly, we have*

$$\mathbf{a} \geq 0, \quad K\mathbf{a} \geq 0, \quad \sum_{i=0}^{n} A_i = N, \quad \textit{and } A_0 = 1.$$

Here K is an $(n+1) \times (n+1)$ matrix with (i,j) entry $K_i(j)$ for $0 \leq i, j \leq n$. Also, $\mathbf{a} \geq 0$ means coordinate wise inequality.

We let $|\mathbf{a}| = N$ denote the size *of the quasicode, and define the* dual *of a quasicode to be the quasicode*

$$\mathbf{a}^{\perp} \triangleq \frac{1}{|\mathbf{a}|} K\mathbf{a}.$$

If $\mathcal{C} \in F_q^n$ is a code, its distance distribution $\mathbf{a}$ is a quasicode with $|\mathbf{a}| = |\mathcal{C}|$. Also, if $\mathcal{C}$ is a linear code, its dual $\mathcal{C}^{\perp}$ has distance distribution $\mathbf{a}^{\perp}$.

As we are studying subcubes, we first prove that the dual of a subcube is a subcube.

Lemma 9. *Let $\mathbf{a} = (A_0, \ldots, A_n)$ where $A_i = \binom{d}{i}(q-1)^i$. Then*

$$\mathbf{a}^{\perp} = (A_0^{\perp}, \ldots, A_n^{\perp}), \quad A_i^{\perp} = \binom{n-d}{i}(q-1)^i.$$

Proof of Lemma 9. Suppose $\mathbb{F}_q$ is a field. There the dual of $\mathcal{C}_{q,d}$ is $\mathcal{C}_{q,n-d}$; this follows from the orthogonal direct sum decomposition $\mathbb{F}_q^n = \mathbb{F}_q^d \oplus \mathbb{F}_q^{n-d}$ with $\mathbb{F}_q^d \perp \mathbb{F}_q^{n-d}$. Moreover, the distance distribution $\mathbf{a}$ of $\mathcal{C}_{q,d}$ is $(A_0, \ldots, A_n)$ with $A_i = \binom{d}{i}(q-1)^i$. Now if q is not a prime power, we are still fine since the dual distance distribution of a subcube is a rational function of finite degree in q given by some combination of Krawtchouk polynomials, and we have an infinite number of q where the expressions agree. Alternatively, one may do a direct computation. □

We now turn to examining the cone of monotonically increasing functions $\mathcal{F}$, and provide a spanning set for it.

Lemma 10. *$\mathcal{F}$ is the nonnegative span of $f_0, f_1, \ldots, f_n$ where $f_j(x) = \mathbb{1}(x \geq j)$.*

Proof of Lemma 10. f_j are obviously monotonically increasing. Moreover, any monotonically increasing f may be written as

$$f(x) = f(0)f_0(x) + \sum_{j=1}^{n}(f(j) - f(j-1))f_j(x).$$

□

We note that the analogous characterization for the cone of completely monotonic functions $\mathcal{G}$ is already provided by [29, Lemma 4].

A slightly more delicate issue is what happens to $\mathcal{F}$ under duality. $\mathcal{G}$ behaves very well, and turns out to be invariant under duality [29, Lemma 10]. $\mathcal{F}$ is unfortunately not invariant under duality. In order to get useful information about the dual of $\mathcal{F}$, it is helpful to develop some recurrence relations for Krawtchouk polynomials:

Lemma 11.

$$K_j(i; n, q) = K_j(i-1; n-1, q) - K_{j-1}(i-1; n-1, q). \tag{2.24a}$$

$$K_j(i; n, q) = K_j(i; n-1, q) + (q-1)K_{j-1}(i; n-1, q). \tag{2.24b}$$

Proof of Lemma 11. Basically (2.24) follows from Definition 7 by applying Pascal's rule in two different ways. We note that (2.24a) is widely known in coding theory (see e.g [118, 1.2.15]).

Explicitly, we have

$$\begin{aligned}
K_j(i; n, q) &= \sum_{k=0}^{j}(-1)^k(q-1)^{j-k}\binom{i}{k}\binom{n-i}{j-k} \\
&= \sum_{k=0}^{j}(-1)^k(q-1)^{j-k}\left(\binom{i-1}{k} + \binom{i-1}{k-1}\right)\binom{n-i}{j-k} \\
&= \sum_{k=0}^{j}(-1)^k(q-1)^{j-k}\binom{i-1}{k}\binom{(n-1)-(i-1)}{j-k} \\
&- \sum_{k=0}^{j}(-1)^{k-1}(q-1)^{(j-1)-(k-1)}\binom{i-1}{k-1}\binom{(n-1)-(i-1)}{(j-1)-(k-1)} \\
&= K_j(i-1; n-1, q) - K_{j-1}(i-1; n-1, q).
\end{aligned}$$

This proves (2.24a).

Using Pascal's rule on the other binomial coefficient in the expansion, we have

$$\begin{aligned}K_j(i;n,q) &= \sum_{k=0}^{j}(-1)^k(q-1)^{j-k}\binom{i}{k}\binom{n-i}{j-k}\\ &= \sum_{k=0}^{j}(-1)^k(q-1)^{j-k}\binom{i}{k}\left(\binom{(n-1)-i}{j-k}+\binom{(n-1)-i}{(j-1)-k}\right)\\ &= K_j(i;n-1,q)+(q-1)K_{j-1}(i;n-1,q).\end{aligned}$$

This proves (2.24b). □

With the recurrence relation (2.24a) in hand, we may characterize the dual of $\mathcal{F}$ as follows.

Lemma 12. *Let $\mathbf{f}_j$ denote the column vector (of length $n+1$) corresponding to $f_j(x) = \mathbb{1}(x \geq j)$, where $0 \leq j \leq n$. Suppose also that $n > 1$. Then,*

$$K^t\mathbf{f}_j =$$
$$\left(\sum_{k=j}^{n}\binom{n}{k}(q-1)^k, -K_{j-1}(0;n-1), -K_{j-1}(1;n-1), \dots, -K_{j-1}(n-1;n-1)\right).$$

We denote the set of functions lying in the nonnegative span of $K^t\mathbf{f}_j$ the dual cone *of the cone of monotonic functions $\mathcal{F}$. Symbolically, we denote this cone by $\mathcal{F}^\perp$.*

Proof of Lemma 12. If $i = 0$, we get

$$\sum_{k=j}^{n}K_k(0) = \sum_{k=j}^{n}\binom{n}{k}(q-1)^k.$$

If $i > 0$, we have

$$\begin{aligned}\sum_{k=j}^{n}K_k(i) &= -\sum_{k=0}^{j-1}K_k(i)\\ &= -K_{j-1}(i-1;n-1,q),\end{aligned}$$

where we used orthogonality followed by telescoping (2.24a). □

There are a number of essentially equivalent ways of proceeding with the proof of Theorem 2, which differ chiefly in how one views dual certificates of optimality in the LP. For example, in the proof of Theorem 1 we worked explicitly with the dual variables p_i. A more conceptually satisfying approach that we shall adopt now is to use the idea of positive definite functions. Specifically we shall use [29, Prop 5].

The construction of dual certificates shall rest on two key inequalities involving Krawtchouk polynomials. The first of these holds for general reasons. Our proof of the second uses (2.24b) to fuel an induction argument.

Lemma 13. *Suppose $0 \leq i, j \leq n$. Then we have the following inequalities involving Krawtchouk polynomials*

$$\forall q \geq 2 \quad K_j(0) \geq K_j(i). \tag{2.25}$$

If $q > 2$, we have

$$\frac{q-1}{q} K_j(1) + \frac{1}{q} K_j(0) + \frac{(-1)^i (K_j(0) - K_j(1))}{q(q-1)^{i-1}} \geq K_j(i). \tag{2.26}$$

Proof of Lemma 13. By positive-definiteness of Krawtchouk polynomials, we immediately conclude (2.25).

For (2.26), we shall induct on n and use (2.24b) to do the induction step. The base case $n = 1$ is trivial since we have equality in (2.26) for $i = 0, 1$ regardless of n. We also note that for any n the case $j = 0$ is trivial as it simply asserts $1 \geq 1$.

Using

$$K_j(0) = \binom{n}{j} (q-1)^j$$

and

$$K_j(1) = \binom{n}{j} (q-1)^j - q(q-1)^{j-1} \binom{n-1}{j-1}$$

we may rewrite (2.26) as

$$K_j(i; n) \leq \binom{n-1}{j} (q-1)^j + (-1)^i (q-1)^{j-i} \binom{n-1}{j-1}. \tag{2.27}$$

Now suppose (2.27) is true for $n-1$ with $n-1 \geq 1$. We wish to establish it for $0 \leq i, j \leq n$. If $j = n$ we have equality in (2.27) since $K_n(i;n) = (-1)^i(q-1)^{n-i}$. If $0 < i, j \leq n-1$, we may use the induction hypothesis to obtain

$$K_j(i;n-1) \leq \binom{n-2}{j}(q-1)^j + (-1)^i(q-1)^{j-i}\binom{n-2}{j-1}, \tag{2.28}$$

$$(q-1)K_{j-1}(i;n-1) \leq \binom{n-2}{j-1}(q-1)^j + (-1)^i(q-1)^{j-i}\binom{n-2}{j-2}. \tag{2.29}$$

Adding (2.28) and (2.29) to each other and using (2.24b) we get

$$\begin{aligned} &K_j(i;n) \\ &\leq \left(\binom{n-2}{j-1} + \binom{n-2}{j}\right)(q-1)^j + (-1)^i(q-1)^{j-i}\left(\binom{n-2}{j-1} + \binom{n-2}{j-2}\right) \\ &= \binom{n-1}{j}(q-1)^j + (-1)^i(q-1)^{j-i}\binom{n-1}{j-1}, \end{aligned}$$

thereby proving the induction step as long as $i \leq n-1$.

All that remains is establishing (2.27) for $i = n$ and $0 < j \leq n-1$. Using

$$K_j(n;n) = (-1)^j\binom{n}{j}$$

and cancelling common factors, we reduce our task to establishing

$$(-1)^j n \leq (n-j)(q-1)^j + (-1)^n j(q-1)^{j-n}. \tag{2.30}$$

If j is odd, it suffices to prove that

$$(q-1)^n \geq \frac{j}{n-j}.$$

Since we have the well known $2^n \geq n-1$, we get

$$(q-1)^n \geq 2^n \geq n-1 \geq \frac{j}{n-j}$$

as $j \leq n-1$, thus proving the j odd case.

Now suppose j is even. If n is even, it suffices to prove

$$\frac{n}{n-j} \leq (q-1)^j.$$

But we have

$$\frac{n}{n-j} \leq j+1 \leq 2^j \leq (q-1)^j$$

using the well known $2^j \geq j+1$. Now suppose n is odd. We note that it suffices to prove (2.30) for $n = j+1$, since for a fixed j incrementing n by 1 increases the left hand side by 1, while the right hand side increases by at least 4 since $q \geq 3$. Now for $n \geq 3$ we have by an obvious induction

$$\frac{3n}{2} - \frac{1}{2} \leq 2^{n-1}.$$

Thus, we have

$$n + \frac{n-1}{q-1} \leq n + \frac{n-1}{2} \leq 2^{n-1} \leq q^{n-1},$$

completing the proof of (2.30). □

Lemma 13 will allow us to prove Theorem 2 quite easily. We use the framework of [29, Prop 5].

Proof of Theorem 2. We begin by noting that the statement for the full space (2.5) is trivial. Slightly less trivial, but also clear is (2.2).

Nevertheless, we shall examine (2.2) from the LP perspective, in order to then use duality and obtain (2.4).

As we are working with the subcube of cardinality q, we already have quite a lot of information on what to use for h. Specifically, we know $h(i) = c_0 + c_n K_n(i)$ for some $c_n \geq 0$. We need $h(1) = f(1)$, this gives us

$$c_0 - c_n(q-1)^{n-1} = f(1).$$

A simple choice is to use $c_0 = f(1)$ and $c_n = 0$. This choice works as long as

$$f(1) \leq f(k) \; \forall k > 1, \tag{2.31}$$

yielding the desired objective value of $(q-1)f(1)$. Constraint (2.31) is met by all $f \in \mathcal{F}$, thus proving (2.2) from the LP perspective, although it is trivial to see directly as well. More importantly we note that by (2.25) and Lemma 12 all functions in $\mathcal{F}^\perp$ also satisfy (2.31). By a simple modification of [29, Prop 9] to use $\mathcal{F}$ and $\mathcal{F}^\perp$ instead of $\mathcal{G}$, we have proved the $\mathcal{C}_{q,n-1}$ case (2.4).

We now turn to the cube of size q^2. Duality will then allow us to obtain the corresponding statements for q^{n-2}. Once again, we know that

$$h(i) = c_0 + c_n K_n(i) + c_{n-1} K_{n-1}(i)$$

for some $c_n, c_{n-1} \geq 0$ and that we also need $h(1) = f(1)$ and $h(2) = f(2)$. A simple choice is to use $c_{n-1} = 0$. Thus, we have $h(i) = c_0 + c_n(-1)^i(q-1)^{n-i}$. Solving the simultaneous linear equations $h(1) = f(1)$ and $h(2) = f(2)$ for c_0, c_n, we get

$$h(i) = f(1)\frac{1}{q} + f(2)\frac{q-1}{q} + \frac{(-1)^i(f(2)-f(1))}{q(q-1)^{i-2}}. \tag{2.32}$$

We note that such an h, if it satisfies $h(i) \leq f(i) \; \forall 1 \leq i \leq n$, yields the desired objective value since

$$\begin{aligned}
q^2 c_0 - h(0) &= qf(1) + q(q-1)f(2) - c_0 - c_n(q-1)^n \\
&= qf(1) + q(q-1)f(2) - c_0 - \frac{(q-1)^2(f(2)-f(1))}{q} \\
&= \left(q - \frac{1}{q} + \frac{(q-1)^2}{q}\right) f(1) + \left(q(q-1) - \frac{q-1}{q} - \frac{(q-1)^2}{q}\right) f(2) \\
&= 2(q-1)f(1) + (q-1)^2 f(2).
\end{aligned}$$

It is clear from the form of h (2.32) that $h(i)$ lies in between $f(1)$ and $f(2)$ for all $1 \leq i \leq n$, regardless of the value of q. Now, if f is a monotonically increasing function, such an h is a valid dual certificate. In particular, any $f \in \mathcal{F}$ satisfies this constraint, so we have proved (2.3). As $\mathcal{G} \subseteq \mathcal{F}$ and $\mathcal{G}$ is invariant under duality, by [29, Prop 9] we have proved (2.7). In fact, we can conclude the analog of (2.7) for general q. Nonetheless, we wish to prove something stronger for $q > 2$, namely (2.6).

Here (2.26) plays an important role. First, note that by the analog of [29, Prop 9] for $\mathcal{F}, \mathcal{F}^{\perp}$ it suffices to show universal optimality over $\mathcal{F}^{\perp}$ for $\mathcal{C}_{q,2}$. Equivalently we may show it over a spanning set of $\mathcal{F}^{\perp}$, without loss that given by 12.

The validity of h given by (2.32) would follow from the inequality

$$-\frac{1}{q}K_{j-1}(0) - \frac{q-1}{q}K_{j-1}(1) + \frac{(-1)^i(-K_{j-1}(1) + K_{j-1}(0))}{q(q-1)^{i-2}} \leq -K_{j-1}(i-1). \quad (2.33)$$

Negating and relabelling the indices $j-1 \to j, i-1 \to i$, and the implicit index $n-1 \to n$, (2.33) is nothing but (2.26). This completes the proof of (2.6), and in turn Theorem 2. □

2.5 Connection to maximum noise stability

The main goal of this section is to elaborate upon the connection between the "ground state" problem for anticodes and "maximal noise stability". The key idea is that an anticode may be viewed as the inverse image of $\{0\}$ (or by complementarity $\{1\}$) of a Boolean valued function. Furthermore, it is not unreasonable to expect a connection, since we know that by the edge isoperimetric inequality, subcubes are optimal sets in terms of their edge boundary. Thus the edge isoperimetric inequality implies that subcubes are optimal anticodes with respect to a potential function that is zero everywhere except at distance 1, where it is negative. The corresponding functions, namely the $\mathbf{and}_k$ functions have very good noise stability when k is small, especially in the low noise limit. In fact, the derivative of noise stability with respect to the noise parameter at noise level 0 is an affine function of the size of the edge boundary [89, Prop 2.51].

We first prove a lemma that makes this connection precise. As this lemma does not necessarily require a product noise structure, we derive it in slightly greater generality.

Lemma 14. *Let $f : \mathbb{F}_q^n \to \{0,1\}$ be a Boolean valued function, and let $\mathcal{C} = f^{-1}(\{0\})$. Let $s(\cdot|\cdot)$ be a row-stochastic $q^n \times q^n$ transition probability matrix. We assume that s has an additive noise structure that is equidistributed over Hamming shells. In other words, s is determined by the vector $(s_0, \ldots, s_n)$ with $s(y|x) = s_{|x-y|}$, $s_i \geq 0$, and $\sum_{i=0}^n \binom{n}{i}(q-1)^i s_i = 1$. Let $h(i) = s_i$ be a potential function. Then, we have:*

$$\mathbf{Stab}_s(n,f) = 1 + 2|\mathcal{C}|q^{-n}(E_h(\mathcal{C}) + s_0 - 1). \tag{2.34}$$

Proof of Lemma 14.

$$\begin{aligned}
\mathbf{Stab}_s(n,f) &= \Pr(f(\mathbf{x}) = f(\mathbf{y})) \\
&= \frac{1}{q^n}\left(\sum_{x\in\mathcal{C}} \Pr(\mathbf{y}\in\mathcal{C}|\mathbf{x}=x) + \sum_{x\in\mathcal{C}^c} \Pr(\mathbf{y}\in\mathcal{C}^c|\mathbf{x}=x)\right) \\
&= \frac{1}{q^n}\left(\sum_{x\in\mathcal{C}} \Pr(\mathbf{y}\in\mathcal{C}|\mathbf{x}=x) + q^n - |\mathcal{C}| - \sum_{x\in\mathcal{C}^c} \Pr(\mathbf{y}\in\mathcal{C}|\mathbf{x}=x)\right) \\
&= \frac{1}{q^n}\left(\sum_{x\in\mathcal{C}} \Pr(\mathbf{y}\in\mathcal{C}|\mathbf{x}=x) + q^n - |\mathcal{C}| - \sum_{x\in\mathcal{C}} \Pr(\mathbf{y}\in\mathcal{C}^c|\mathbf{x}=x)\right) \\
&= \frac{1}{q^n}\left(\sum_{x\in\mathcal{C}} \Pr(\mathbf{y}\in\mathcal{C}|\mathbf{x}=x) + q^n - |\mathcal{C}| - |\mathcal{C}| + \sum_{x\in\mathcal{C}} \Pr(\mathbf{y}\in\mathcal{C}|\mathbf{x}=x)\right) \\
&= \frac{1}{q^n}\left(2\sum_{x\in\mathcal{C}} \Pr(\mathbf{y}\in\mathcal{C}|\mathbf{x}=x) + q^n - 2|\mathcal{C}|\right) \qquad (2.35) \\
&= 1 - 2|\mathcal{C}|q^{-n} + 2q^{-n}\left(\sum_{x\in\mathcal{C},y\in\mathcal{C}} s_{|x-y|}\right) \\
&= 1 + 2|\mathcal{C}|q^{-n}(E_h(\mathcal{C}) + s_0 - 1).
\end{aligned}$$

□

We have chosen to highlight (2.35) as we will need this intermediate step later in this chapter. This step depends upon the symmetry $s(y|x) = s(x|y)$, but does not need additive noise structure. We remark that $\mathbf{Stab}_s(n,f) = \mathbf{Stab}_s(n,\overline{f})$ together with the above proof imply the "particle-antiparticle" relation involving the potential energy [29, Sec VII].

In an application to noise stability, one naturally specializes the above Lemma 14 to the BSC and q-SC families. This gives us Corollary 1 and Corollary 2.

Proof of Corollaries 1 and 2. For general q, we have for the q-SC

$$\begin{aligned} s_i &= (1-\epsilon)^{n-i}\left(\frac{\epsilon}{q-1}\right)^i \\ &= (1-\epsilon)^n\left(\frac{\epsilon}{(q-1)(1-\epsilon)}\right)^i \end{aligned}$$

Observe that if

$$0 \le \epsilon \le 1 - \frac{1}{q},$$

we have

$$0 \le \frac{\epsilon}{(q-1)(1-\epsilon)} \le 1.$$

Thus s gives a completely monotonic potential h in Lemma 14. Lemma 14 shows that maximizing noise stability subject to a cardinality constraint is equivalent to maximizing $E_h(\mathcal{C})$. We may then use Theorem 2 (specifically eqs. (2.4), (2.6) and (2.7)) together with complementarity ($\mathbf{Stab}_s(n, f) = \mathbf{Stab}_s(n, \overline{f})$) to obtain Corollary 1 and Corollary 2. □

We remark that [29, Prop 29] allows us to remove an arbitrary point of a subcube of measure $1/(q^2)$ and get universal optimality for a set of measure $1/(q^2) - 1/(q^n)$. By complementarity, one can add an arbitrary point to the complement of a subcube to get universal optimality for a set of measure $(q^2-1)/(q^2)+1/(q^n)$. Similar remarks apply to the other subcubes. We did not explicitly record these facts in Corollary 1 and Corollary 2 as such operations result in an asymptotically vanishing perturbation of measure. Similarly, we did not record explicitly the noise stability analog for $\mathcal{C}_{q,2}$ or $\mathcal{C}_{q,1}$ since these subcubes have an asymptotically vanishing measure.

We also find it illustrative to express the noise stability across the q-SC of an anticode in terms of its dual distance distribution. One application of this is in providing a quick way to deduce the intuitive fact that the noise stability is monotonically non-increasing from $\epsilon = 0$ to $\epsilon = 1 - 1/q$ (Corollary 4), something which is unclear from the expression (2.34). We note that we are essentially rephrasing the discussion of [89, Sec. 2.4] in slightly different language and for general q; see also [89, Ex. 5.28]. The link to Fourier analysis on the hypercube should not be too mysterious to the reader, especially in view of our preliminary remarks 1.

Proposition 4. *Let $\epsilon \in [0, 1]$. Define the* correlation factor

$$\rho \triangleq 1 - \frac{q\epsilon}{q-1}.$$

Then we have

$$\mathbf{Stab}_\epsilon(n, \mathcal{C}) = 1 + 2\mu \left(\mu \sum_{k=0}^{n} \rho^k A_k^\perp - 1 \right). \tag{2.36}$$

Here, $A_k^\perp$ denotes the dual distance distribution of $\mathcal{C}$.

Proof. By the generating function for Krawtchouk polynomials (2.13), we have

$$\begin{aligned} h(i) &= (1-\epsilon)^n \left(\frac{\epsilon}{(q-1)(1-\epsilon)} \right)^i \\ &= q^{-n} (1 + (q-1)\rho)^{n-i} (1-\rho)^i \\ &= q^{-n} \sum_{k=0}^{n} K_k(i; n, q) \rho^k. \end{aligned}$$

Then we have by (2.34)

$$\begin{aligned} \mathbf{Stab}_\epsilon(n, f) &= 1 + 2\mu \left(\sum_{i=0}^{n} A_i h(i) - 1 \right) \\ &= 1 + 2\mu \left(q^{-n} \sum_{k=0}^{n} \rho^k \sum_{i=0}^{n} A_i K_k(i; n, q) - 1 \right) \\ &= 1 + 2\mu \left(\mu \sum_{k=0}^{n} \rho^k A_k^\perp - 1 \right). \end{aligned}$$

□

As an immediate corollary of 4, we have the desired monotonicity in ϵ of the noise stability.

Corollary 4. $\mathbf{Stab}_\epsilon(n, \mathcal{C})$ *is monotonically decreasing in* ϵ *for* $0 \leq \epsilon \leq 1 - \frac{1}{q}$. *Furthermore, with* $\mu = \frac{|\mathcal{C}|}{q^n}$, *we have*

$$\mathbf{Stab}_\epsilon(n, \mathcal{C}) \geq \mu^2 + (1-\mu)^2.$$

Proof of Corollary 4. The interval for ϵ corresponds precisely to $0 \leq \rho \leq 1$. The lower bound follows since $\rho = 0$ corresponds to $\mathbf{x} \perp \mathbf{y}$. □

2.6 A mean value theorem for noise stability

Although the BSC and q-SC are entirely reasonable channel models for noise stability that are perhaps most useful for current applications, one may wonder what can be said about maximal noise stability with respect to other channels. In general, this is not an easy question. For example, the LP bounds (or the SDP generalizations) will work only for a noise that respects the underlying symmetries of Hamming space as seen in Lemma 14. In particular, it is easy to see that the only product noise that respects such symmetries is the one given by the q-SC.

Intuitively, it seems clear that for a general channel maximal noise stability should be at least as high as that for a q-SC with noise level being chosen appropriately to match the "average" chance of a bit flip. This is because one should be able to "tailor" a function better for a channel that does not treat coordinates and letters symmetrically. One route to proving such things is by formulating a random coding/mean value theorem for noise stability. Making these intuitive ideas precise is the subject of Theorem 3.

Proof of Theorem 3. First, note that t is nonnegative and row-stochastic, and thus a valid probability kernel, by its definition (2.8). This ensures that (2.9) is well defined. We have the following

$$\begin{aligned}
&\frac{1}{|\mathrm{Aut}|}\sum_{\sigma\in\mathrm{Aut}}\mathbf{Stab}_s(n,\sigma f)\\
&=\frac{1}{|\mathrm{Aut}|}\sum_{\sigma\in\mathrm{Aut}}\Pr((\sigma f)(\mathbf{x})=(\sigma f)(\mathbf{y}))\\
&=\frac{1}{|\mathrm{Aut}|}\sum_{\sigma\in\mathrm{Aut}}\frac{1}{q^n}\sum_{x,y\in\mathbb{F}_q^n}\mathbb{1}(f(\sigma x)=f(\sigma y))s(y|x)\\
&=\frac{1}{q^n}\sum_{x,y\in\mathbb{F}_q^n}\sum_{\sigma\in\mathrm{Aut}}\frac{1}{|\mathrm{Aut}|}\mathbb{1}(f(\sigma x)=f(\sigma y))s(y|x)\\
&=\frac{1}{q^n}\sum_{x,y\in\mathbb{F}_q^n}\sum_{\sigma\in\mathrm{Aut}}\frac{1}{|\mathrm{Aut}|}\mathbb{1}(f(x)=f(y))s(\sigma^{-1}y|\sigma^{-1}x)\\
&=\frac{1}{q^n}\sum_{x,y\in\mathbb{F}_q^n}\mathbb{1}(f(x)=f(y))\left(\frac{1}{|\mathrm{Aut}|}\sum_{\sigma\in\mathrm{Aut}}s(\sigma^{-1}y|\sigma^{-1}x)\right)\\
&=\frac{1}{q^n}\sum_{x,y\in\mathbb{F}_q^n}\mathbb{1}(f(x)=f(y))t(y|x)\\
&=\mathbf{Stab}_t(n,f).
\end{aligned}$$

This completes the proof. □

Using Theorem 3 we can immediately derive the following corollary for maximal noise stability.

Corollary 5. *Using the notation of Theorem 3, we have*

$$\mathbf{Stab}_s^*(n,\mu)\geq\mathbf{Stab}_t^*(n,\mu). \tag{2.37}$$

Proof of Corollary 5. An important feature of the proof of Theorem 3 is that the expected value of a Boolean function is invariant under our chosen group action, that is $\mathbb{E}[f(\mathbf{x})]=\mathbb{E}[(\sigma f)(\mathbf{x})]$. Let a function f be chosen such that subject to $\mathbb{E}[f(\mathbf{x})]=\mu$, f attains maximal noise stability under $t(\cdot|\cdot)$. By the above Theorem 3, we see that there exists a $\sigma\in\mathrm{Aut}$ such that $\mathbf{Stab}_s(n,\sigma f)\geq\mathbf{Stab}_t^*(n,\mu)$. Thus, $\mathbf{Stab}_s^*(n,\mu)\geq\mathbf{Stab}_s(n,\sigma f)\geq\mathbf{Stab}_t^*(n,\mu)$. This completes the proof. □

We may specialize 5 to the case of an i.i.d product kernel $s(\cdot|\cdot)$ to get a "symmetrized" $t(\cdot|\cdot)$ that corresponds to a q-BSC.

Corollary 6. *Let $s(b^n|a^n) = \prod_{i=1}^{n} r(b_i|a_i)\ \forall a^n \in \mathbb{F}_q^n$ be a product kernel, where $r(\cdot|\cdot)$ is a $q \times q$ probability kernel. Let $\epsilon = \frac{q - tr(r)}{q}$, where $tr(r)$ denotes the trace of the kernel r viewed as a $q \times q$ row-stochastic matrix. Then, we have*

$$\mathbf{Stab}_r^*(n, \mu) \geq \mathbf{Stab}_\epsilon^*(n, \mu). \tag{2.38}$$

Proof of Corollary 6. Let Π_q denote the permutation group on q letters. We know that the distance preserving automorphisms of $\mathbb{F}_q^n$ consist of permutations of the n coordinates composed with arbitrary permutations of the individual coordinates. Using this decomposition, by (2.8), we have

$$\begin{aligned} t(y|x) &= \frac{1}{|\mathrm{Aut}|} \sum_{\sigma \in \mathrm{Aut}} s(\sigma y|\sigma x) \\ &= \frac{1}{|\mathrm{Aut}|} \sum_{\sigma \in \mathrm{Aut}} \prod_{i=1}^{n} r((\sigma y)_i|(\sigma x)_i) \\ &= \frac{n!}{|\mathrm{Aut}|} \prod_{i=1}^{n} \sum_{\sigma \in \Pi_q} r(\sigma y_i|\sigma x_i) \\ &= \prod_{i=1}^{n} \left[\frac{1}{q!} \sum_{\sigma \in \Pi_q} r(\sigma y_i|\sigma x_i) \right] \\ &= \left(\frac{\mathrm{tr}(r)}{q} \right)^{n-|x-y|} \left(\frac{q - \mathrm{tr}(r)}{q(q-1)} \right)^{|x-y|}. \end{aligned}$$

This proves (2.38). □

2.7 Large n and balls versus subcubes

One aspect of the anticoding problem that we find intriguing is the role of balls versus subcubes in Hamming space. For example, the optimal anticodes in the *isodiametric* sense given by [7, Diametric Theorem] involve Cartesian products of balls and subcubes. In the context of binary noise stability, at small expected value $\mu \to 0$

and high noise $\epsilon \to 1/2$, it is well known that Hamming balls of appropriate radius maximize noise stability in a sharp sense as $n \to \infty$ [110, Prop. 2.2]. On the other hand, we know thanks to Corollary 1 that for $\mu \in \{1/4, 1/2, 3/4\}$ we maximize noise stability regardless of $\epsilon \leq 1/2$ by using Hamming subcubes and their complements. The goal of this section is to further explore the question of which sets maximize noise stability subject to an expected value constraint.

It is perhaps useful to define a limiting notion of noise stability for a product noise that captures $n \to \infty$. The formulation of a limiting value for noise stability avoids issues such as how many points one needs to pick from the last shell of the Hamming ball in order to meet a cardinality constraint, at the cost of possibly missing out on finite n phenomena. Although we do not make much further use of this limiting notion once we establish it, we find it conceptually satisfying. Moreover, we do make further use in our proof of Theorem 4 of some auxiliary Lemmas 15, 18 that are established en route to this definition.

Once defined rigorously, we shall denote the maximum noise stability for a symmetric kernel r on $\mathcal{X} \times \mathcal{X}$ by $\mathbf{Stab}_r^*(\infty, \mu)$. In an analogous manner to our other notation, we may specialize to the q-SC with parameter ϵ and denote it by $\mathbf{Stab}_\epsilon^*(\infty, \mu)$.

Our approach shall be to define the limit on q-adic μ first, and then extend by uniform continuity to all $\mu \in [0, 1]$. The first task is simple, and so we do it first.

In order to do so, it is helpful to understand the general behavior of Cartesian products of anticodes under a product noise.

Lemma 15. *Let $\mathcal{C}_m \subseteq \mathbb{F}_q^m$ have measure $\mu_m = |\mathcal{C}_m|/q^m$. Similarly, let $\mathcal{C}_n \subseteq \mathbb{F}_q^n$ have measure $\mu_n = |\mathcal{C}_n|/q^n$. Let $r(\cdot|\cdot)$ be a $q \times q$ symmetric transition probability matrix. Let $\mathcal{C} = \mathcal{C}_m \times \mathcal{C}_n$. Then its measure $\mu = |\mathcal{C}|/q^{m+n} = \mu_m \mu_n$, and its noise stability is given by*

$$\mathbf{Stab}_r(m+n, \mathcal{C}) + 2\mu - 1 = \frac{1}{2}(\mathbf{Stab}_r(m, \mathcal{C}_m) + 2\mu_m - 1)(\mathbf{Stab}_r(n, \mathcal{C}_n) + 2\mu_n - 1).$$

Proof of Lemma 15. The measure statement is trivial. The stability statement follows readily from (2.35), as follows.

$$
\begin{aligned}
&\mathbf{Stab}_r(m+n,\mathcal{C}) + 2\mu - 1\\
&= \frac{2}{q^{m+n}} \sum_{x^m \in \mathcal{C}_m} \sum_{x^n \in \mathcal{C}_n} \Pr(\mathbf{y} \in \mathcal{C} | \mathbf{x} = x)\\
&= \frac{2}{q^{m+n}} \sum_{x^m \in \mathcal{C}_m} \sum_{x^n \in \mathcal{C}_n} \Pr(\mathbf{y}^m \in \mathcal{C}_m | \mathbf{x}^m = x^m) \Pr(\mathbf{y}^n \in \mathcal{C}_n | \mathbf{x}^n = x^n)\\
&= \frac{1}{2}\left[\frac{2}{q^m} \sum_{x^m \in \mathcal{C}_m} \Pr(\mathbf{y}^m \in \mathcal{C}_m | \mathbf{x}^m = x^m)\right]\left[\frac{2}{q^n} \sum_{x^n \in \mathcal{C}_n} \Pr(\mathbf{y}^n \in \mathcal{C}_n | \mathbf{x}^n = x^n)\right]\\
&= \frac{1}{2}(\mathbf{Stab}_r(m,\mathcal{C}_m) + 2\mu_m - 1)(\mathbf{Stab}_r(n,\mathcal{C}_n) + 2\mu_n - 1).
\end{aligned}
$$

□

Lemma 16. *Let $\mu = \frac{a}{q^k}$ be a q-adic fraction. Let $r(\cdot|\cdot)$ be a row-stochastic $q \times q$ transition probability matrix. Then the following limit exists and may be used to define*

$$
\mathbf{Stab}_r^*(\infty,\mu) \triangleq \lim_{n \to \infty} \mathbf{Stab}_r^*(n,\mu). \tag{2.39}
$$

Proof of Lemma 16. The sequence at hand is uniformly bounded by 1. Furthermore, we claim that it is non-decreasing. Let $\mathcal{A}_m^*$ denote an optimal set for noise stability at μ for $n = m$. Consider $\mathcal{A}_{m+1} = \mathcal{A}_m^* \times \mathbb{F}_q$. Then by Lemma 15,

$$
\mathbf{Stab}_r^*(m+1,\mu) \geq \mathbf{Stab}_r(m+1,\mathcal{A}_{m+1}) = \mathbf{Stab}_r(m,\mathcal{A}_m^*) = \mathbf{Stab}_r^*(m,\mu).
$$

□

The slightly trickier task is to prove uniform continuity. Our approach to this is to use a randomly chosen subset of the appropriate cardinality to get a reasonably good *subanticode* given an optimal anticode. Our approach in fact yields Lipschitz continuity.

The "averaging" step is contained in the following

Lemma 17. *Let $a_{ij}, 1 \leq i,j \leq n$ denote a collection of reals. Let $m \leq n$, and let $\mathcal{A}$*

denote the collection of m-subsets of $[n]$. Then, we have

$$\frac{1}{|\mathcal{A}|}\sum_{\mathcal{B}\in\mathcal{A}}\sum_{(i,j)\in\mathcal{B}\times\mathcal{B}} a_{ij} = \frac{m}{n}\sum_{i=1}^{n} a_{ii} + \frac{m(m-1)}{n(n-1)}\sum_{i\neq j} a_{ij}. \tag{2.40}$$

Furthermore, if a_{ij} are nonnegative, we have the immediate estimate

$$\frac{1}{|\mathcal{A}|}\sum_{\mathcal{B}\in\mathcal{A}}\sum_{(i,j)\in\mathcal{B}\times\mathcal{B}} a_{ij} \geq \frac{m(m-1)}{n(n-1)}\sum_{i,j} a_{ij}. \tag{2.41}$$

Proof of Lemma 17. The fraction of the number of times a given diagonal element appears is $\frac{\binom{m}{1}}{\binom{n}{1}}$. Similarly, for an off-diagonal element, it is $\frac{\binom{m}{2}}{\binom{n}{2}}$. This proves (2.40). The estimate (2.41) follows immediately from $m \leq n$ and $a_{ij} \geq 0$. □

Lemma 17 together with the noise stability expression (2.35) allow one to readily understand the noise stability of random subanticodes.

Lemma 18. *Let $m' \leq m \leq q^n$ denote two cardinalities, and let $\mu' = \frac{m'}{q^n}, \mu = \frac{m}{q^n}$ denote their respective measures. Let $\mathcal{C}$ denote an anticode of size m. Let $\mathcal{A}$ denote the collection of anticodes $\mathcal{C}'$ of size m' obtained as m'-subsets of $\mathcal{C}$. Then*

$$\frac{1}{|\mathcal{A}|}\sum_{\mathcal{C}'\in\mathcal{A}} \mathbf{Stab}_r(n,\mathcal{C}') \geq (1-2\mu') + \frac{\mu'(m'-1)}{\mu(m-1)}\left(\mathbf{Stab}_r(n,\mathcal{C}) + 2\mu - 1\right). \tag{2.42}$$

Proof of Lemma 18.

$$\begin{aligned}\frac{1}{|\mathcal{A}|}\sum_{\mathcal{C}'\in\mathcal{A}} \mathbf{Stab}_r(n,\mathcal{C}') &= (1-2\mu') + \frac{2}{q^n}\left(\frac{1}{|\mathcal{A}|}\sum_{\mathcal{C}'\in\mathcal{A}}\sum_{(x,y)\in\mathcal{C}'\times\mathcal{C}'} \Pr(\mathbf{y}=y|\mathbf{x}=x)\right)\\ &\geq (1-2\mu') + \frac{m'(m'-1)}{m(m-1)}\left(\frac{2}{q^n}\sum_{(x,y)\in\mathcal{C}\times\mathcal{C}} \Pr(\mathbf{y}=y|\mathbf{x}=x)\right)\\ &= (1-2\mu') + \frac{\mu'(m'-1)}{\mu(m-1)}\left(\mathbf{Stab}_r(n,\mathcal{C}) + 2\mu - 1\right).\end{aligned}$$

□

The above Lemmas 16, 18 allow us to uniquely define $\mathbf{Stab}_r^*(\infty,\mu)$ via

Proposition 5. *There exists a unique Lipschitz continuous (in μ) extension of* $\mathbf{Stab}_r^*(\infty, \mu)$ *defined on a dense subset via* (2.39) *to all* $\mu \in [0, 1]$.

Proof of Proposition 5. Let $\mu' = m'/q^k, \mu = m/q^k$ be two q-adic measures in $(0, 1)$. Without loss of generality suppose $\mu' \leq \mu$. Let $\epsilon > 0$, and choose $n \geq k$ large enough such that we have simultaneously

$$\mathbf{Stab}_r^*(\infty, \mu') - \mathbf{Stab}_r^*(n, \mu') \leq \epsilon, \tag{2.43}$$

$$\mathbf{Stab}_r^*(\infty, \mu) - \mathbf{Stab}_r^*(n, \mu) \leq \epsilon, \tag{2.44}$$

$$\mathbf{Stab}_r^*(\infty, 1-\mu') - \mathbf{Stab}_r^*(n, 1-\mu') \leq \epsilon, \tag{2.45}$$

$$\mathbf{Stab}_r^*(\infty, 1-\mu) - \mathbf{Stab}_r^*(n, 1-\mu) \leq \epsilon, \tag{2.46}$$

$$\frac{\mu'^2}{\mu^2} - \frac{\mu'(m'-1)}{\mu(m-1)} \leq \epsilon,$$

$$\frac{(1-\mu)^2}{(1-\mu')^2} - \frac{(1-\mu)(q^n-m-1)}{(1-\mu')(q^n-m'-1)} \leq \epsilon.$$

We note that estimates (2.43), (2.44) are ineffective, as we do not know how fast we converge to the large n limit, though Lemma 16 guarantees that we get there eventually. The following two (2.45), (2.46) follow from (2.43), (2.44) by taking complements, and the remainder are effective.

Then by Lemma 18, we have a lower bound on $\mathbf{Stab}_r^*(\infty, \mu') - \mathbf{Stab}_r^*(\infty, \mu)$

$$\begin{aligned}
&\mathbf{Stab}_r^*(\infty, \mu') - \mathbf{Stab}_r^*(\infty, \mu) \\
&\geq -2\epsilon + (1-2\mu') + \left(\frac{\mu'(m'-1)}{\mu(m-1)} - 1\right)\mathbf{Stab}_r^*(n, \mu) + \frac{\mu'(m'-1)}{\mu(m-1)}(2\mu-1) \\
&\geq -2\epsilon + (1-2\mu') + \left(\frac{\mu'(m'-1)}{\mu(m-1)} - 1\right) + \frac{\mu'(m'-1)}{\mu(m-1)}(2\mu-1) \\
&\geq -2\epsilon + (1-2\mu') + \left(\frac{\mu'^2}{\mu^2} - 1 - \epsilon\right) + \frac{\mu'^2}{\mu^2}(2\mu-1) - \epsilon|2\mu-1| \\
&\geq -4\epsilon + (1-2\mu') + \left(\frac{\mu'^2}{\mu^2} - 1\right) + \frac{\mu'^2}{\mu^2}(2\mu-1) \\
&= -4\epsilon - \frac{2\mu'(\mu-\mu')}{\mu} \\
&\geq -4\epsilon - 2(\mu-\mu').
\end{aligned} \tag{2.47}$$

Using the complementarity relations $\mathbf{Stab}_r^*(n, \mu) = \mathbf{Stab}_r^*(n, 1-\mu)$ and their limiting equivalents $\mathbf{Stab}_r^*(\infty, \mu) = \mathbf{Stab}_r^*(\infty, 1-\mu)$ (valid at this stage for q-adic μ), we may get an upper bound in similar manner to (2.47) as

$$\begin{aligned} \mathbf{Stab}_r^*(\infty, \mu') - \mathbf{Stab}_r^*(\infty, \mu) &= -(\mathbf{Stab}_r^*(\infty, 1-\mu) - \mathbf{Stab}_r^*(\infty, 1-\mu')) \\ &\leq 4\epsilon + 2(\mu - \mu'). \end{aligned} \tag{2.48}$$

As $\epsilon > 0$ was arbitrary, combining eqs. (2.47) and (2.48) gives the estimate

$$|\mathbf{Stab}_r^*(\infty, \mu') - \mathbf{Stab}_r^*(\infty, \mu)| \leq 2|\mu - \mu'|. \tag{2.49}$$

The uniform continuity on the dense subset given by the q-adic fractions then extends uniquely to a uniformly continuous function of all $\mu \in [0, 1]$ (see e.g. [100, Prob. 4.13]), and in fact a Lipschitz continuous function with Lipschitz constant 2. □

Remark 3. *The constant* 2 *is the best possible in the estimate* (2.49)*. For example, consider* $\epsilon = (q-1)/q$ *and* r *the q-SC(*ϵ*). Then,* $\mathbf{y} \perp \mathbf{x}$*, and so* $\mathbf{Stab}_r^*(\infty, \mu) = \mu^2 + (1-\mu)^2$*. Differentiating at* $\mu = 0$ *demonstrates the tightness of* (2.49)*.*

We also note that $\mathbf{Stab}_r^*(\infty, \mu)$ *is not necessarily differentiable everywhere. Take for instance* $r(y|x) = \mathbb{1}(y = 0)$*. Then,*

$$\mathbf{Stab}_r^*(\infty, \mu) = \mathbf{Stab}_r^*(n, \mu) = \max(\mu, 1-\mu).$$

Our goal now is to explore the role of balls versus subcubes in the context of noise stability for the q-SC in more detail. First, we give an example where balls do better than subcubes for high values of noise. We have the following

Proposition 6. *For* $q \geq 3$ *and* $n \geq q^2+q+1$*,* $\mathcal{C}_{q,3}$ *is not universally optimal for noise stability across the family of q-SC(*ϵ*),* $0 \leq \epsilon \leq 1 - 1/q$*. More specifically, for such n,* $\mathcal{C}_{q,3}$ *is not optimal for noise stability in the interval* $((2q+1)/(1+q)^2, 1-1/q)$*. For* $q = 2$*, for all* $n \geq 3$*,* $\mathcal{C}_{2,3}$ *is universally optimal. For* $q = 2$*,* $n \geq 12$*,* $\mathcal{C}_{2,4}$ *is not universally optimal. More specifically, for such n,* $\mathcal{C}_{2,4}$ *is not optimal for noise* *stability in the interval* $\big((19-\sqrt{137})/16, 1/2\big) \supset (0.456, 0.5)$.

Proof of Proposition 6. As in Lemma 9, we know that the distance distribution of $\mathcal{C}_{q,k}$ is A where $A_i = \binom{k}{i}(q-1)^i$. Thus its energy with respect to

$$h(i) = (1-\epsilon)^n \left(\frac{\epsilon}{(q-1)(1-\epsilon)}\right)^i$$

is

$$\begin{aligned} E_h(\mathcal{C}_{q,k}) &= (1-\epsilon)^n \sum_{i=1}^{k} \binom{k}{i} (q-1)^i \left(\frac{\epsilon}{(q-1)(1-\epsilon)}\right)^i \\ &= (1-\epsilon)^n((1-\epsilon)^{-k} - 1). \qquad (2.50) \end{aligned}$$

Now suppose $q \geq 3$, and consider the Hamming ball of radius 1, denoted by $\mathcal{B}_{q^2+q+1,q,1}$. The cardinality of this ball is $1 + n(q-1) = q^3$. We remark that this is the smallest cube we could hope for by Theorem 2. The distance distribution of this Hamming ball is B where $B_0 = 1, B_1 = (q^3-1)/q^2, B_2 = (q^3-1)(q^2-1)/q^2$ and $B_i = 0$ for $i > 2$. Thus

$$\begin{aligned} E_h(\mathcal{B}_{q^2+q+1,q,1}) &= (1-\epsilon)^n \left[\frac{\epsilon(q^3-1)}{(1-\epsilon)(q-1)q^2} + \frac{\epsilon^2(q^3-1)(q^2-1)}{(1-\epsilon)^2(q-1)^2q^2}\right] \\ &= (1-\epsilon)^n \left[\frac{\epsilon(q^2+q+1)(\epsilon q+1)}{(1-\epsilon)^2 q^2}\right]. \qquad (2.51) \end{aligned}$$

Thus for $k = 3$, we see by eqs. (2.50) and (2.51) that $E_h(\mathcal{B}_{q^2+q+1,q,1}) \geq E_h(\mathcal{C}_{q,3})$ precisely when

$$\frac{(q^2+q+1)(\epsilon q+1)(1-\epsilon)}{q^2} \geq \epsilon^2 - 3\epsilon + 3. \qquad (2.52)$$

Treating (2.52) as a quadratic in ϵ, we see that the roots of this quadratic are

$$r_1 = 1 - \frac{1}{q}, \quad r_2 = \frac{2q+1}{(1+q)^2}.$$

Note that r_1 is expected as it corresponds to $\mathbf{y} \perp \mathbf{x}$, in which case the noise stability does not depend on the actual anticode.

We claim that for $q \geq 3$, $r_2 < r_1$. Cross multiplying, this reduces to showing that for $q \geq 3$,

$$f(q) = q^3 - q^2 - 2q - 1 > 0.$$

This claim may be proved readily. For example, at $q = 3$ the inequality is true as $11 > 0$. Differentiating, we get $f'(q) = 3q^2 - 2q - 2$, and the roots of $f'(q) = 0$ are $(1 \pm \sqrt{7})/3$, which are both less than 3. This root check completes the proof of the $q \geq 3$ case, since for $n > q^2 + q + 1$, we may simply take a Cartesian product with $\{0\}$ on $n - (q^2 + q + 1)$ coordinates.

We now turn to $q = 2$. Here we can not rely on using $\mathcal{C}_{2,3}$, since in fact it may easily be checked by hand (using Proposition 3 to simplify the case analysis) that $\mathcal{C}_{2,3}$ is universally optimal for $n \geq 3$. Hence we move to $\mathcal{C}_{2,4}$, and use an "almost-Hamming ball" for $n = 12$ and of cardinality 16. This time, we take care of the last shell by filling it in lexicographic order. The distance distribution B of this anticode $\mathcal{B}$ is given by $B_0 = 1, B_1 = 9/4, B_2 = 9, B_3 = 15/4$ and $B_i = 0$ for $i > 3$. Then

$$E_h(\mathcal{B}) = (1-\epsilon)^n \left[\frac{9\epsilon}{4(1-\epsilon)} + \frac{9\epsilon^2}{(1-\epsilon)^2} + \frac{15\epsilon^3}{4(1-\epsilon)^3} \right]. \tag{2.53}$$

Thus by eqs. (2.50) and (2.53), we see that $E_h(\mathcal{B}) \geq E_h(\mathcal{C}_{2,4})$ precisely when

$$\frac{4\epsilon - 6\epsilon^2 + 4\epsilon^3 - \epsilon^4}{(1-\epsilon)^4} \leq \frac{3\epsilon(3 + 6\epsilon - 4\epsilon^2)}{4(1-\epsilon)^3}.$$

Cross multiplying, this boils down to studying when

$$g(x) \triangleq 16x^3 - 46x^2 + 33x - 7 \geq 0.$$

As we already know one root $s_1 = 1/2$, we may easily find the other roots

$$\{s_2, s_3\} = \frac{19 \pm \sqrt{137}}{16}.$$

The root below $1/2$ is what matters for us. Once again, the derivative checks out, so we have our desired interval. This completes the proof. □

We remark that by no means is Proposition 6 tight in the sense of the obtained measures. The parameters were chosen above to reflect the smallest cube cardinalities where universal optimality does not exist. As can be seen in the proof above, it also has the advantage of producing a low degree polynomial that can be analyzed by hand readily. For example, the above proof yields for $q = 2$ an anticode example of measure $1/256$. It turns out one can do far better:

Remark 4. *Let $n = 19$ and $q = 2$. Then, $\mathcal{C}_{2,14}$ is not universally optimal across the BSC-ϵ family. More specifically, $\mathcal{C}_{2,14}$ is not optimal for $\epsilon \in (0.484, 0.5)$. Note that the measure here is $1/32$. The example is simply an "almost-Hamming ball" of the appropriate cardinality, with the last shell filled in lexicographic order.*

We note that the lack of universal optimality at measure $1/32$, or more broadly Proposition 6, is not mysterious, and may be understood more conceptually as follows. The discussion here follows closely [89, Sec 5.4], and we give a summary. The derivative of noise stability with respect to ϵ at high noise $\epsilon = 1/2$ is proportional to the *degree-1 Fourier weight*. For a Hamming ball, as $n \to \infty$, the *degree-1 Fourier weight* as a function of the measure is given by the square of the Gaussian isoperimetric function due to the central limit theorem ([89, Propn. 5.25]). The corresponding quantity for subcubes is also given in [89, pg 125]. One may then numerically compute the values for measure $1/32$ and compare. This consideration tells us that for some large enough n, one should be able to construct an "almost-Hamming ball" of measure $1/32$ that does better than the corresponding cube for high noise. The above Remark 4 is simply a numerical quantification; $n = 19$ was the smallest n for which the "almost-Hamming ball" happened to work.

We also note that $1/32$ represents the largest measure where this phenonmenon occurs; at $1/16, 1/8$ one can check that subcubes do better than balls at high noise, and for $1/4, 1/2$ we have universal optimality of subcubes by Theorem 2. We suspect that there is universal optimality for measures $1/16, 1/8$, and raise the following

Conjecture 1. *Let $q = 2$. Then for all $n \geq 4$, $\mathcal{C}_{2,n-3}, \mathcal{C}_{2,n-4}$ are universally optimal with respect to the cone of all negations of completely monotonic functions $\mathcal{G}$ (defined* *in Theorem 2).*

We have the following numerical evidence in favor of the 1/8 case of Conjecture 1. One may use the "order-1" SDP bounds of [102] to study this problem (we used SDPA-GMP for this purpose, see e.g. [85]); they represent a natural generalization of the LP bounds. These bounds do manage to certify universal optimality for the 1/8 case for $n \leq 8$, but unfortunately do not do so for $n = 9$ onwards. For the 1/16 case, there seems to be no nontrivial certificates: even $n = 7$ is not certified, even though we know that $\mathcal{C}_{2,3}$ is universally optimal (see e.g. Proposition 6). It is possible that a higher constant order SDP bound will certify universal optimality for all n, at least for 1/8.

We shall now use Proposition 6 to show that "universal optima are sparse for anticoding".

Our approach to showing this is by understanding how anticodes behave in Hamming space when they are *stacked.* We accordingly have

Lemma 19. *Let $c = kq^l$, $c' = r$ be given, where $c + c' \leq q^n$, $0 < c' < q^l$. Let $\mu = (c + c')/q^n, \mu' = c'/q^l$. Let $\mathcal{C}_1, \mathcal{C}_2 \subset \mathbb{F}_q^l$ be two anticodes of cardinality c'. Let*

$$\mathcal{D}_1 = \left(\{\overline{k}_q\} \times \mathcal{C}_1\right), \mathcal{D}_2 = \left(\{\overline{k}_q\} \times \mathcal{C}_2\right),$$

live in $\mathbb{F}_q^n$. Let $\mathcal{B} = \mathbf{L}(\mathbb{F}_q^n, c)$. Consider the anticodes $\mathcal{A}_1, \mathcal{A}_2$ obtained by disjoint union

$$\mathcal{A}_1 = \mathcal{B} \,\dot\cup\, \mathcal{D}_1,$$

$$\mathcal{A}_2 = \mathcal{B} \,\dot\cup\, \mathcal{D}_2.$$

We say that $\mathcal{A}_1, \mathcal{A}_2$ are obtained by stacking. *Then we have*

$$\mathbf{Stab}_\epsilon(n, \mathcal{A}_1) - \mathbf{Stab}_\epsilon(n, \mathcal{A}_2) = \frac{\mu(1-\epsilon)^{n-l}}{\mu'}(\mathbf{Stab}_\epsilon(l, \mathcal{C}_1) - \mathbf{Stab}_\epsilon(l, \mathcal{C}_2)). \quad (2.54)$$

Proof of Lemma 19. We observe that the only differences between the distance distributions of $\mathcal{A}_1$ and $\mathcal{A}_2$ come from the distances internal to $\mathcal{D}_1, \mathcal{D}_2$. This is because $\mathcal{B}$ is common to both, and the interactions between $\mathcal{D}_1, \mathcal{D}_2$ and $\mathcal{B}$ are identical due to the symmetric nature of the projection of $\mathcal{B}$ onto the lower l coordinates. As such, letting

$$h(i) = (1-\epsilon)^n \left(\frac{\epsilon}{(q-1)(1-\epsilon)}\right)^i, \quad h'(i) = (1-\epsilon)^l \left(\frac{\epsilon}{(q-1)(1-\epsilon)}\right)^i,$$

we have by Lemma 14

$$\begin{aligned}
\mathbf{Stab}_\epsilon(n, \mathcal{A}_1) - \mathbf{Stab}_\epsilon(n, \mathcal{A}_2) &= 2\mu(E_h(\mathcal{A}_1) - E_h(\mathcal{A}_2)) \\
&= 2\mu(E_h(\mathcal{D}_1) - E_h(\mathcal{D}_2)) \\
&= 2\mu(E_h(\mathcal{D}_1) - E_h(\mathcal{D}_2)) \\
&= 2\mu(1-\epsilon)^{n-l}(E_h'(\mathcal{D}_1) - E_h'(\mathcal{D}_2)) \\
&= \frac{\mu(1-\epsilon)^{n-l}}{\mu'}(\mathbf{Stab}_\epsilon(l, \mathcal{C}_1) - \mathbf{Stab}_\epsilon(l, \mathcal{C}_2)).
\end{aligned}$$

□

With Lemma 19, we may deduce the lack of universal optimality at cardinality $c + c'$ for $\mathbb{F}_q^n$ from the lack of universal optimality at c' for $\mathbb{F}_q^l$. Thus combining with Proposition 6 already gives us many more cardinalities where we lack universal optimality than the examples furnished by Proposition 6 itself. Nevertheless, we may do much better by combining further with Lemmas 15, 18. In particular, this combination suffices to achieve our aim here, namely that "universal optima are sparse for anticoding".

We have all the ingredients in place to prove Theorem 4, except for a technicality that involves estimating the difference in noise stability between lex-sets of different sizes. This technicality may be viewed as analogous to Lemma 18, except that this time the anticodes at hand are nicely structured. In fact, this Lemma 20 only depends on the nesting of two anticodes $\mathcal{C} \subseteq \mathcal{C}'$.

Lemma 20. *Let $\mathcal{C} \subseteq \mathcal{C}' \subseteq \mathbb{F}_q^n$ be two anticodes of measures μ, μ' respectively, and* *consider noise stability with respect to the q-SC(ϵ). Then we have the estimate*

$$|\mathbf{Stab}_\epsilon(n, \mathcal{C}') - \mathbf{Stab}_\epsilon(n, \mathcal{C})| \leq 4(\mu' - \mu).$$

Proof of Lemma 20. As usual, we define $h(i) = (1-\epsilon)^n \left(\frac{\epsilon}{(q-1)(1-\epsilon)}\right)^i$. Observe that each codeword in $\mathcal{C}' \setminus \mathcal{C}$ can have at most $\binom{n}{i}(q-1)^i$ codewords at Hamming distance i from it. From this observation, we can bound the difference of energy with respect to h and hence the noise stability (by Lemma 14) by

$$\begin{aligned}
\mathbf{Stab}_\epsilon(n, \mathcal{C}') - \mathbf{Stab}_\epsilon(n, \mathcal{C}) &\leq 2(\mu' - \mu) + 2q^{-n} \sum_{i=0}^{n} h(i) |\mathcal{C}' \setminus \mathcal{C}| \binom{n}{i} (q-1)^i \\
&= 2(\mu' - \mu) + 2(1-\epsilon)^n (\mu' - \mu) \sum_{i=0}^{n} \left(\frac{\epsilon}{1-\epsilon}\right)^i \binom{n}{i} \\
&= 4(\mu' - \mu). \qquad (2.55)
\end{aligned}$$

We may apply complements as in the proof of Proposition 5 to get the symmetric version of (2.55), namely

$$\begin{aligned}
\mathbf{Stab}_\epsilon(n, \mathcal{C}') - \mathbf{Stab}_\epsilon(n, \mathcal{C}) &= -(\mathbf{Stab}_\epsilon(n, \overline{\mathcal{C}}) - \mathbf{Stab}_\epsilon(n, \overline{\mathcal{C}'})) \\
&\geq 4(\mu - \mu').
\end{aligned}$$

This completes the proof. □

All the pieces are now in play to prove Theorem 4. We remark that all estimates we use here are effective unlike Proposition 5 since we focus on finite but large enough n, though we prefer not to quantify the required parameters for simplicity.

Proof of Theorem 4. Proposition 6 guarantees the existence of a cube size $k(q)$, a suitably large $n_0(q)$, an anticode $\mathcal{C}(q)$, an interval $I(q) = (\epsilon(q), 1 - 1/q)$, and a lower bound $c(q) > 0$ such that for all $\epsilon \in I(q)$,

$$\mathbf{Stab}_\epsilon(n_0(q), \mathcal{C}(q)) - \mathbf{Stab}_\epsilon(n_0(q), \mathcal{C}_{q,k(q)}) \geq c(q). \qquad (2.56)$$

For $n' \geq n_0(q)$, considering the product anticode

$$\mathcal{D}(n', s, q) = \mathcal{C}(q) \times \mathbf{L}(\mathbb{F}_q^{n'-n_0(q)}, s)$$

for any $1 \leq s \leq q^{n'-n_0(q)}$, we get by Lemma 15 and (2.56),

$$\begin{aligned}&\mathbf{Stab}_\epsilon(n', \mathcal{D}(n', s, q)) - \mathbf{Stab}_\epsilon(n', \mathbf{L}(\mathbb{F}_q^{n'}, sq^{k(q)})) = \\ &\frac{c(q)}{2}(\mathbf{Stab}_\epsilon(n' - n_0(q), \mathbf{L}(\mathbb{F}_q^{n'-n_0(q)}, s)) + 2sq^{n_0(q)-n'} - 1).\end{aligned} \tag{2.57}$$

We can lower bound the right hand side of (2.57) by Corollary 4 to get

$$\mathbf{Stab}_\epsilon(n', \mathcal{D}(n', s, q)) - \mathbf{Stab}_\epsilon(n', \mathbf{L}(\mathbb{F}_q^{n'}, sq^{k(q)})) \geq c(q)\mu'(q, s, n')^2, \tag{2.58}$$

where $\mu'(q, s, n) = s/q^{n'-n_0(q)}$. The goal is a uniform lower bound, so we specialize to s an integer satisfying

$$q^{n'-n_0(q)-n_1(q)} \leq s \leq q^{n'-n_0(q)}.$$

Letting

$$\mu(q) \triangleq q^{k(q)-n_0(q)}, \nu(q) = q^{k(q)-n_0(q)-n_1(q)},$$

we have by the above (2.58) a uniform lower bound (call it $c'(q)$) for measures

$$\nu(q), \nu(q) + q^{k(q)-n'}, \nu(q) + 2q^{k(q)-n'}, \ldots, \mu(q).$$

We now interpolate to the full range of integers in $[\nu(q), \mu(q)]$ by using the above measures as anchors. We do this by using a random subanticode for the first term on the left of (2.58) (cf. Lemma 18), and the estimate for the difference of noise stability between lex-sets for the second term (cf. Lemma 20). As $\nu(q), \mu(q)$ are fixed and bounded away from 0, there exists an $n_2(q)$ such that for any $n' \geq n_2(q)$ and cardinality $r \in [\nu(q)q^{n'}, \mu(q)q^{n'}]$, there exists a subanticode $\mathcal{D}'(n', r, q)$ with

$$|\mathbf{Stab}_\epsilon(\mathcal{D}'(n', r, q)) - \mathbf{Stab}_\epsilon(\mathcal{D}(n', s(r), q))| < \frac{c'(q)}{3}, \tag{2.59a}$$

$$|\mathbf{Stab}_\epsilon(n', \mathbf{L}(\mathbb{F}_q^{n'}, s(r)q^{k(q)})) - \mathbf{Stab}_\epsilon(n', \mathbf{L}(\mathbb{F}_q^{n'}, r))| < \frac{c'(q)}{3}. \tag{2.59b}$$

Here $s(r)$ denotes the largest integer of the form $sq^{k(q)}$ at or below r. The estimates eqs. (2.58) and (2.59) yield

$$\mathbf{Stab}_\epsilon(n', \mathcal{D}'(n', r, q)) - \mathbf{Stab}_\epsilon(n', \mathbf{L}(\mathbb{F}_q^{n'}, r)) > \frac{c'(q)}{3}$$
$$\forall r \in [\nu(q)q^{n'}, \mu(q)q^{n'}], n' \geq n_2(q), \epsilon \in I(q). \tag{2.60}$$

We may now stack $\mathcal{D}'(n', r, q)$ on top of $\mathcal{B} = \mathbf{L}(\mathbb{F}_q^n, r')$ to get an anticode $\mathcal{A}(n, r+r', q)$ that does better than $\mathbf{L}(\mathbb{F}_q^n, r+r')$ for $\epsilon \in I(q)$ as long as the conditions of Lemma 19 are met.

One simple set of sufficient conditions on the cardinality r and size n is the following. Let $r = \sum_{i=0}^{n-1} r_i q^{n-i}$ be the base q expansion of r. Suppose we ignore the last $n_2(q)$ digits, and focus on $(r_0, r_1, \ldots, r_{n-n_2(q)-1})$. Let $a(q) \leq b(q)$ be two integers such that $1 + \log_q(\mu(q)^{-1}) \leq a(q) \leq b(q) \leq -1 + \log_q(\nu(q)^{-1})$. Note that this is always possible as we had flexibility in the choice of $\nu(q)$, indeed we can make $[a(q), b(q)]$ have arbitrarily long, but constant (independent of n) length.

Suppose that there is a run of zeros in $(r_0, \ldots, r_{n-n_2(q)-1})$ of length l with $a(q) \leq l \leq b(q)$. Then, there is an anticode $\mathcal{A} \subseteq \mathbb{F}_q^n$ with $|\mathcal{A}| = r$ that does better than $\mathbf{L}(\mathbb{F}_q^n, r)$ for noise stability in the range $I(q)$. In particular, for these cardinalities we do not have universal optimality of lex-sets, which are the only candidates for universal optimality. It is also clear that the set of such cardinalities $\mathcal{T}$ is contained in $\overline{\mathcal{S}}$, and that it also has measure $|\mathcal{T}|/q^n \geq 1 - o\left(c''(q)^n\right)$, for some $c''(q) < 1$, simply because we have to avoid a finite pattern. This completes the proof. □

2.8 Open problems

Many problems here remain open. What we find most attractive is the one we raise in Conjecture 1, as it would complete the classification of which subcubes are universally optimal for noise stability. One can also ask analogous questions for general q.

We also lay out a far more general version of this problem, where we ask for characterizing the sharp value of

$$s_q^*(\mu, \epsilon) \triangleq \mathbf{Stab}_\epsilon^*(\infty, \mu), \quad (\mu, \epsilon) \in \left[0, \frac{1}{2}\right] \times \left[0, 1 - \frac{1}{q}\right].$$

Other μ may be obtained by complementing, and other ϵ no longer reflect an anti-coding problem. We give our speculations regarding the $q = 2$ case based on what we understood from proving Theorem 4. Let us consider a dyadic $\mu = 0.a_1a_2 \dots a_k$, where the a_k denote bits in a binary expansion. By the stacking construction used in proving Theorem 4, we know that with a sufficiently long run of zeros in the a_i, and large enough ϵ, we can stack some anticode on top of a lex-set to do better than a lex-set of the same augmented cardinality in terms of noise stability. The anticodes we used in the proof were bootstrapped from a finite phenomenon(e.g. Proposition 6, Remark 4), with the finite phenomenon being given by an almost-Hamming ball. It is perhaps more natural to stack almost-Hamming balls directly, and in fact it may be possible to prove Theorem 4 by such an approach. However, it does not seem easy to express the noise stability of a Hamming ball with measure different from 1/2 in convenient form, with even the degree-1 Fourier weight (that describes the limit $\epsilon \to 1/2$) involving the Gaussian isoperimetric function, and the full noise stability expression involving the Gaussian quadrant probability (see e.g. [89, Ex. 5.32]. On the other hand, it is possible that the focus on $n = \infty$ helps with some technical issues, such as the fact that we don't know of a closed form for the distance distribution of a Hamming ball, even for $q = 2$.

As our focus was on a finite n statement, for the proof, we favored the above bootstrapping approach. Nevertheless, it is possible that examining the set bits of μ's binary expansion, and using the most favorable stacking of an almost-Hamming ball out of them on top of a lex-set will yield $s_2^*(\mu, \epsilon)$. We note that the relevant distance distributions and noise stability may be computed in polynomial time (in n) for a fixed cardinality. With numerical simulation, we were unable to come up with any better construction of anticodes for large n. We note that our "stacking" proposal here is to some extent anticipated by the reviewer comments described in [50]. Basically, the proposal of [50] was to simply use lex-sets without taking into account the ball/subcube interplay that was pointed out by the reviewer. Thus much about $s_q^*(\mu, \epsilon)$ remains open; indeed Theorem 2 characterizes just a couple of lines in the (μ, ϵ) square, while Theorem 4 shows that for a.s μ, we lack universally optimal anticodes.

Lastly, we find the use of Fourier analytic techniques for isodiametry intriguing, and wish to return to this theme in future work. We also refer the reader to more general remarks in our concluding Chapter 5 regarding the use of Fourier analytic techniques for geometric questions.

Chapter 3

Near-Optimal Coded Apertures for Imaging via Nazarov's Theorem

Given a convex set $B \in \mathbb{R}^2$ (or $\mathbb{R}^n$, for Bang's solution it is all the same), is it possible to cover it by several strips of total width less than the width of B ? (The width of a convex body is defined as the width of the narrowest strip containing it).
The conjectured answer was "no", but it took about 40 years to prove that. The proof, when found, was . . . 2 pages long and required from the reader only basic knowlege of elementary geometry. For reader's convenience it is included into this paper as an appendix.
Of course, to solve the coefficient problem as it was stated above you cannot just apply the result (. . .), but it turns out that a minor modification of the proof is enough. (So minor that I actually even do not pretend to be an author of the next two sections; rather I act there like a shadow that enters and goes over many strange places which completely eliminate the attention of his master just passing by).

Fedor Nazarov, 1997

3.1 Introduction

In Chapter 2, we saw the role that Fourier analysis on Hamming space played in the resolution of a question in theoretical computer science. Along the way we also developed some non Fourier-analytic machinery towards answering natural follow-up questions, such as:

1. What happens for channels that are not BSC(ϵ)?
2. Can we completely classify the universally optimal anticodes in Hamming space?

In this chapter, we return to Fourier analysis, this time on finite cyclic groups $\mathbb{Z}/n\mathbb{Z}$, or in terms more familiar for engineers, the DFT (discrete Fourier transform). To fix notation and our choice of normalization, let us first define for convenience $e(z) \triangleq e^{2\pi i z}$ as is commonly done in say analytic number theory. We then define the DFT of a length n vector $\vec{a} = (a_0, a_1, \ldots, a_{n-1})$ to be a length n vector $\widehat{\vec{a}}$ given by

$$\widehat{\vec{a}}_k = \sum_{j=0}^{n} \vec{a}_j e\left(-\frac{jk}{n}\right).$$

It is also convenient for us to write

$$\vec{1} \triangleq (1, 1, \ldots, 1).$$

Some illustrative examples to make sure that we are on the same page regarding the choice of normalization are

$$\vec{a} = (n, 0, 0, \ldots, 0) \Leftrightarrow \widehat{\vec{a}} = (n, n, \ldots, n),$$
$$\vec{b} = (1, 1, 1, \ldots, 1) \Leftrightarrow \widehat{\vec{b}} = (n, 0, \ldots, 0).$$

With this notational clarification, let us examine a problem that arises in the context of computational imaging. Certain modern imaging systems, especially those operating at high frequencies, use coded apertures. In these systems, a spatial mask that selectively blocks light from reaching the sensor is used as opposed to a traditional

Figure 3-1: “De Radio Astronomica et Geometrica”, Gemma Frisius, 1545

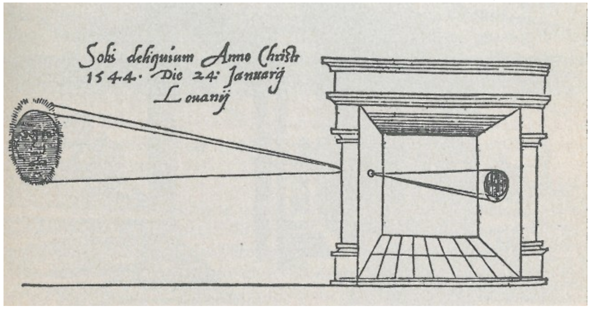


Figure 3-2: Proposal of [37]

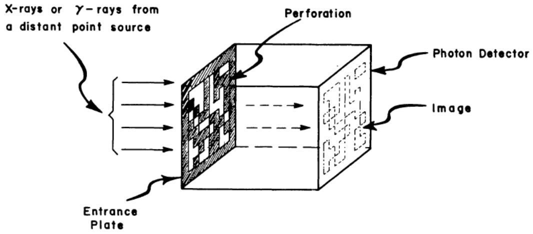


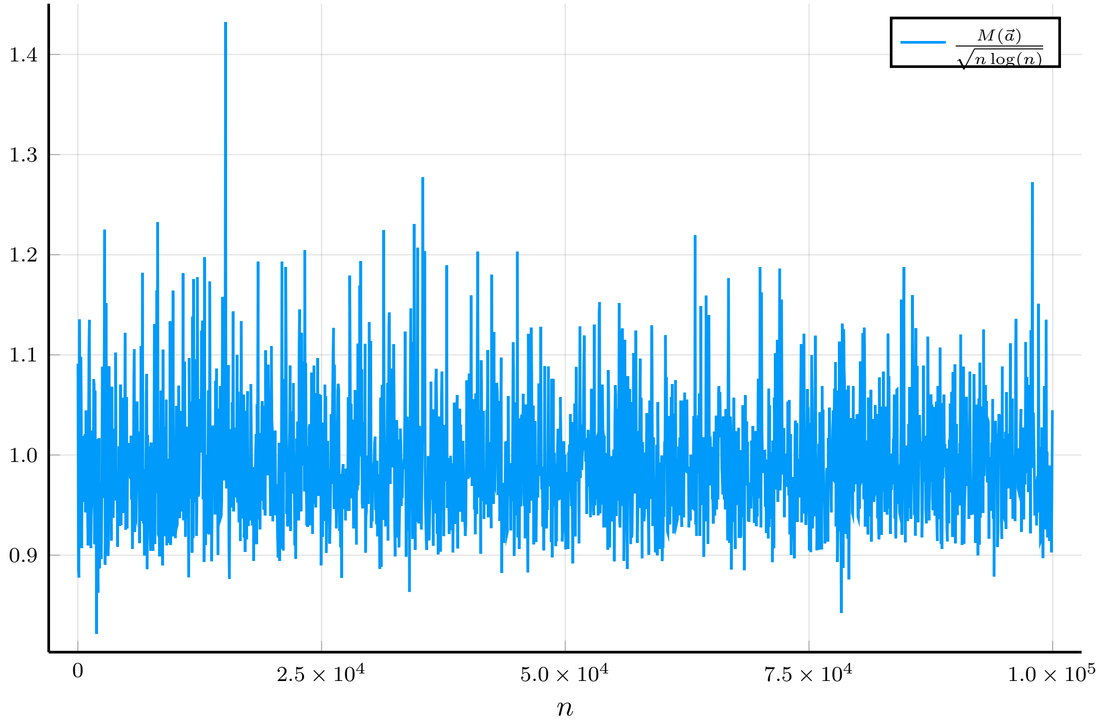
FIG. 1.—Scatter-hole camera. The entrance plate is randomly perforated to provide randomly positioned pinholes. The image is recorded photographically, photoelectrically, or with a photon counter, such as a wire spark chamber.

lens. The scene is then recovered by suitable post-processing. Perhaps the earliest and simplest instance of coded aperture imaging is the pinhole structure (c. 470-391 BCE, Mozi in China [87, pp. 97-99]); see, e.g., [126] for a survey and Figure 3-1 for an illustration.

The development of X-ray and gamma-ray astronomy gave rise to more sophisticated coded apertures [1, 37] to get around the lack of lenses and mirrors in such settings. For the reader curious as to why it can be difficult to fabricate lenses and mirrors for such high frequencies, it turns out that most materials end up having a refractive index near 1 at such frequencies. This fact may be understood from standard dipole/induced dipole and associated oscillator analysis, and an elementary treatment may be found in e.g. [45, Ch. 32].

Both Ables and Dicke [1, 37] proposed using random blockage patterns with a specified mean transmittance as a method to increase the aperture size as compared to the classical pinhole while retaining its resolution benefits. Dicke’s construction is shown in Figure 3-2. Naturally, this also requires a certain nontrivial decoding procedure to recover the image from the superposition formed from the various copies

Figure 3-3: Maximum spectral magnitude of random on-off, normalized

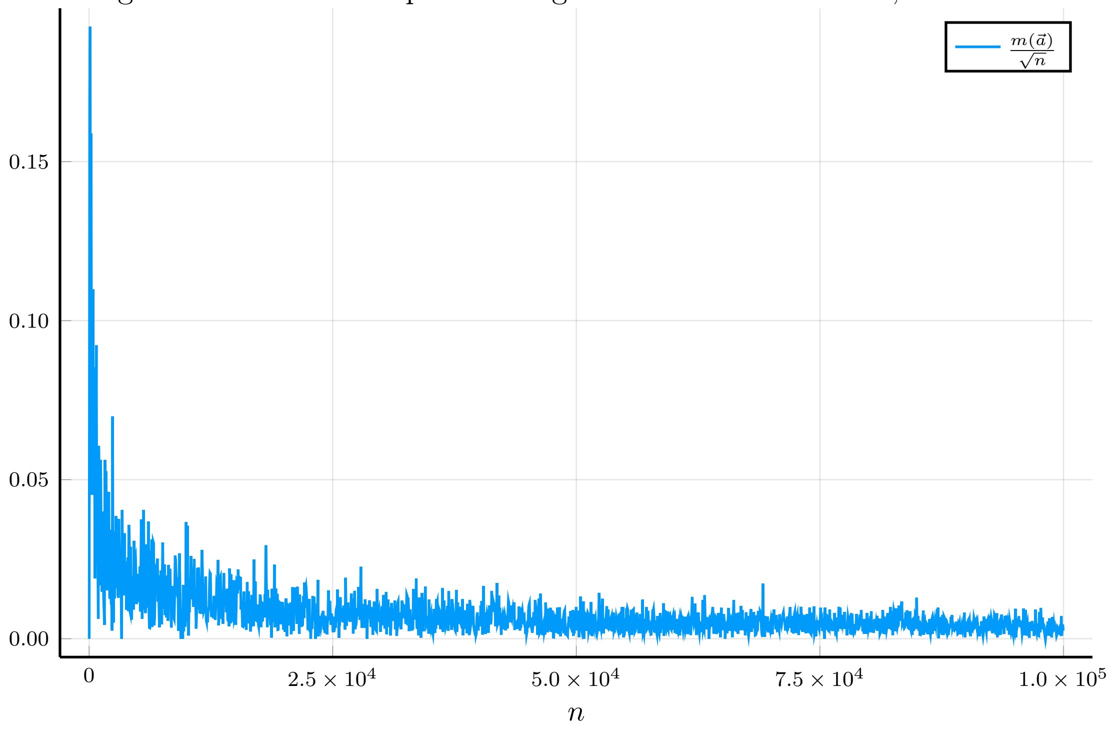


across the different holes. The decoding may done (classically) in analog, or digitally with a computer. In particular, Dicke [37, Fig 2-4] proposed various beautiful analog decoders.

A more modern development of special importance to this chapter is the usage of uniformly redundant arrays (URA) to improve upon random on-off patterns [44]. At a high level, the reason for their superior performance is that the DFT of such a pattern $\vec{a}$ is "spectrally flat", that is $|\widehat{\vec{a}}_i|$ is constant across all $i \neq 0$. "Spectral flatness" is not even close to being achieved with a random on-off pattern drawn from the i.i.d. Bernoulli(0.5) ensemble across the n components, as can be verified numerically by plotting e.g. $M(\vec{a}) \triangleq \max_{i\neq 0} |\widehat{\vec{a}}_i|$ for $i \neq 0$ versus $m(\vec{a}) \triangleq \min_{i\neq 0} |\widehat{\vec{a}}_i|$ for a random on-off pattern (a random $(0, 1, 1, 0, 0, 0, \ldots, 1)$ vector of length n). For a spectrally flat pattern with $\sum_i \vec{a}_i = \Theta(n)$, $M(\vec{a}), m(\vec{a}) = \Theta(\sqrt{n})$. As can be seen in Figures 3-3, 3-4, for a random on-off pattern, this is certainly not the case.

From a historical perspective, this mathematical phenomenon was studied in great depth by Salem and Zygmund in their paper on trigonometric sums with random

Figure 3-4: Minimum spectral magnitude of random on-off, normalized



signs [101, Ch. 4]. In particular, [101, Ch. 4] elucidates the $M(\vec{a}) = \Theta(\sqrt{n \log(n)})$ behavior observed in the plot.

Other modern developments include anti-pinhole imaging [25], as well as the combining of mask and lens in order to, e.g., facilitate depth estimation [75], deblur out-of-focus elements in an image [131], enable motion deblurring [95], and/or recover 4D lightfields [119]. Even more recent work seeks to forgo lenses altogether to decrease costs and meet physical constraints [40, 11]. Understanding coded apertures is also relevant in non-line-of-sight applications where masks naturally occur as scene occlusions [115, 113].

In light of the increased importance of coded apertures, prior work [125] described a model under which they can be analyzed. This model uses far-field geometric optics to model light propagation and a sensor model that includes thermal and shot noise components. Our analysis [8] is performed with respect to a slight refinement of the model of [125]. The basic element at play is an aperture, which we model as a discrete vector of length n. Our choice of a discrete vector reflects a one-dimensional imaging model, and is done purely for notational convenience and clarity. We request patience from the reader who does not find such a restriction reasonable, since all will become clear eventually. See in particular Sec. 3.4 for remarks on this aspect.

We denote the aperture by $\vec{a} \in \mathbb{R}^n$, and we assume it satisfies $\vec{a} \geq 0$ entry-wise. Basically, the aperture entries model how much incoming power is sent onwards. For example, a larger entry means that more light is let through at that location of the aperture, and an entry of zero means that all light is blocked. Thus, a key notion that both [8, 125] account for is that of *transmissivity.*

Definition 14. *For an aperture $\vec{a}$, we define its* transmissivity $\rho(\vec{a})$ *by*

$$\rho(\vec{a}) \triangleq \frac{1}{n} \sum_i \vec{a}_i.$$

Together with mutual information (MI) as a performance metric, [125] compared the classical random on-off apertures [1, 37] of varying transmissivity (think of i.i.d. Bernoulli(p) ensemble to target a transmissivity p) to the "spectrally flat" patterns with transmissivity $1/2$ (same as the URA of [44])[1]. Among other things, the analysis showed that when shot noise dominates thermal noise, randomly generated masks with lower transmissivity than $1/2$ offered greater performance compared to spectrally flat patterns of transmissivity $1/2$.

Our work here extends the work of [125] in multiple respects that may be broadly grouped into the following three main contributions.

First, we refine the prior model of [125] by incorporating exposure time. The effects of exposure time are illustrated well in the original proposal of [37], and accordingly the model of [44] (based off the PhD thesis [23]) incorporates it. Our model is essentially equivalent to that of [44]. At a high level, a lengthier exposure time clearly improves the signal to noise ratio (SNR), though the exact correspondence between the two depends on the sources of noise (thermal vs shot noise) and their statistics. However, this has associated costs, such as the inability to capture motion

[1] Technically, the transmissivity of the URA is $1/2 + O\left(1/n\right)$. We omit this extra term in subsequent nontechnical discourse.

accurately. For background on these sources of noise and how they affect reconstruction, we recommend [44] and the references therein.

Our high level goal is to design apertures that minimize the exposure time needed for a given target reconstruction quality. There are a variety of reconstruction quality metrics one can use, and unfortunately it is usually the case that the most perceptually meaningful metrics are not convenient for mathematical analysis. For example, in the very popular context of video coding, a good state of the art metric is “Video Multimethod Assessment Fusion” (VMAF) developed by Netflix (see e.g. `https://github.com/Netflix/vmaf` and references therein), while a far more convenient metric used classically and in more theoretical literature is peak signal-to-noise-ratio (PSNR). As our focus is on mathematical development, we analyze analytically convenient metrics, specifically mean squared error (MSE). Even that is not enough for our purposes. For example, it is sensitive to the noise model, and for many noise models is quite intractable. We therefore analyze linearly-constrained minimum mean square error (LMMSE) estimation. Separate from considerations of analytical convenience, we defend this choice since most practical decoders (analog/digital) for coded apertures use linear procedures. We note that the prior model of [125] used MI under a Gaussian noise model as a quality measure. At a high level, such a measure may also be written in terms of the spectrum of the aperture, and our analysis is sufficiently general to cover such a measure as well. Broadly and imprecisely, our methods extend to a vast range of quality measures, as long as they can be written in terms of the “content” of the aperture with respect to an orthonormal basis. We elaborate upon these aspects in Sec. 3.4.

Second, we note that the classical URA based apertures have transmissivity $1/2$. Although the spectrum of such apertures is flat in magnitude, a fixed transmissivity performs suboptimally under varying shot noise (proportional to the transmissivity) versus thermal noise levels. As such, we describe how one can construct spectrally flat sequences with transmissivities $1/8, 1/4$ in addition to $1/2$. The spectrally flat sequences with these transmissivities $1/8, 1/4$, when combined with the classical URA work, extend the range of parameters where we have a sharp characterization of optimal coded apertures in our framework. Furthermore, these sequences allow us to obtain a tight answer to the problem of optimal coded apertures for i.i.d. scenes; see Props. 8, 9 for precise statements. Broadly, what we mean by "tight" is in the sense of being within a constant multiplicative factor of optimal for the exposure time. Equivalently, one can express this as a guarantee of being within a constant number of dB in SNR of optimality.

Third, we provide optimal (again up to a constant) coded apertures, both in 1D as well as in 2D, applicable for any prior on the spectrum of the scene at hand. The sense of tightness of the optimality is given precisely in Prop. 10. The priors on the scene spectrum include (but are not limited to) the naturally occuring power law [83]($f^{-\gamma}$-prior). Our aperture design naturally varies depending on the choice of prior, and we provide a (heuristically) efficient greedy algorithm for their generation. As in [125], we use a 1D model to simplify the exposition of concepts and results. We emphasize that all of the results of this chapter generalize naturally to the analogous 2D model, whose discussion we defer to Sec. 3.4.

Essentially all the required mathematical results stem from a beautiful and powerful theorem of Nazarov [86, p. 5] combined with classical waterfilling for spectrum allocation. We find it remarkable here firstly for its mathematical generality and elegance, which among other things allows the analysis to carry over to other quality measures of possible interest such as MI. Secondly, Nazarov's theorem is "effective" in the sense of leading to aperture designs that can be computed in a reasonable amount of time by our greedy algorithm.

To the best of our knowledge, our work here represents the first detailed study of Nazarov's theorem in an applied context. However, we do note that [21, pp. 9-11] has identified other applied problems for which Nazarov's theorem provides conceptual clarity and/or solutions.

3.2 Model

We first describe our model, and discuss how it differs from that in [125]. We use the standard Poisson model of classical optics for photon counting, and emphasize its dependence on the exposure time t. Another perspective is that of far-field, incoherent illumination and the study of the intensity/power transfer, see e.g. [67, Chapter 7] for more information and physical justification. The analysis of MI under Poisson models is cumbersome, and even with mean square error (MSE) it is often unclear how to achieve optimal MSE in practice. As such, the standard estimation process is linear; indeed, the work of [1, 37] used correlation decoders. In fact, both [1, 37] give beautiful analog realizations of such a decoder. Accordingly, we emphasize LMMSE. We note that if one used a Gaussian model instead, LMMSE is the same as MSE. Furthermore, under a Gaussian model, MSE is in turn essentially equivalent to MI in the low SNR limit [106, 59]. LMMSE depends purely on first and second moments, so in our mathematical study we do not emphasize the specific Poisson statistics.

Let $\vec{f}$ denote the intensities of the unknown 1D scene of length n of expected total power J. Let $\mathbb{E}[\vec{f}] = (J/n)\vec{1}, \mathbf{Cov}[\vec{f}] = \mathbf{Q}$. We assume $\mathbf{Q}$ is circulant and diagonalized as $\mathbf{Q} = \mathbf{F}_n^* \mathbf{D} \mathbf{F}_n$; $\mathbf{F}_n$ is the unitary discrete Fourier transform (DFT) matrix and $\mathbf{D} = \operatorname{diag}(\vec{d})$. The measurements at the imaging plane are denoted $\vec{y}_j, j \in [n]$ and the $n \times n$ transfer matrix $\mathbf{A}$ models the aperture. We assume its entries all satisfy $0 \leq \mathbf{A}_{ji} \leq 1/n$ to model that the light can not be redirected, and $\sum_j \mathbf{A}_{ji} \leq 1$ to model local conservation of power. An ideal, perfectly focused, lens may be treated in this setup by $\mathbf{A} = \mathbf{I}$, as it redirects light perfectly.

We assume $\mathbf{A}_{ji} = (1/n)\vec{a}_{i-j \pmod n}$ for a $\vec{a} \geq 0$, i.e. $\mathbf{A}$ is circulant. Let $\rho(\vec{a}) = (1/n)\sum_i \vec{a}_i$ be the *transmissivity* of the aperture. The noise component is denoted by $\vec{z}$ and its statistics are given by $\mathbb{E}[\vec{z}] = 0, \mathbf{Cov}[\vec{z}] = (t(W + J\rho)/n)\mathbf{I}$, where W, J correspond to thermal and shot noise respectively, and t is the *exposure time*. With these, our measurement model is then given by $\vec{y}_j = t\sum_i \mathbf{A}_{ji}\vec{f}_i + \vec{z}_j$, which leads to the following expression for the LMMSE of estimating $\vec{f}$ from $\vec{y}$.

$$m(n, t, W, J, \vec{d}, \vec{a}) = \sum_{i=0}^{n-1} \frac{1}{\frac{1}{\vec{d}_i} + \frac{t|\widehat{\vec{a}}_i|^2}{n(W+J\rho(\vec{a}))}}. \tag{3.1}$$

Here, $\widehat{\vec{a}}$ is the DFT of $\vec{a}$.

3.2.1 Background on linear estimation

We note that (3.1) is an entirely routine derivation for a reader well versed in estimation/statistical signal processing. For a general reader, we provide some background on minimum mean squared error estimation and then derive (3.1). Let $\mathbf{x}, \mathbf{y}$ be a pair of vector valued random variables. Let their means be well-defined and denoted by $\overline{\mathbf{x}}, \overline{\mathbf{y}}$ respectively. The goal is to estimate $\mathbf{x}$ from $\mathbf{y}$, and the minimum squared error is achieved by the conditional expectation $\mathbb{E}[\mathbf{x}|\mathbf{y}]$. As we have discussed, this expression is often intractable, and it is often convenient to restrict the form of the estimator. We study linear estimators here, i.e. estimators of the form $\widehat{\mathbf{x}} = W(\mathbf{y} - \overline{\mathbf{y}}) + b$, and we wish to minimize the expected squared error $\mathbb{E}[(\widehat{\mathbf{x}} - \mathbf{x})^T(\widehat{\mathbf{x}} - \mathbf{x})]$.

In the development we shall assume $\mathbf{x}, \mathbf{y}$ have finite second moments as well. This assumption in turn ensures that the covariance matrices $C_{\mathbf{x}} \triangleq \mathbb{E}[(\mathbf{x} - \overline{\mathbf{x}})(\mathbf{x} - \overline{\mathbf{x}})^T], C_{\mathbf{x},\mathbf{y}} \triangleq \mathbb{E}[(\mathbf{x} - \overline{\mathbf{x}})(\mathbf{y} - \overline{\mathbf{y}})^T], C_{\mathbf{y},\mathbf{x}} \triangleq C_{\mathbf{x},\mathbf{y}}^T, C_{\mathbf{y}} \triangleq \mathbb{E}[(\mathbf{y} - \overline{\mathbf{y}})(\mathbf{y} - \overline{\mathbf{y}})^T]$ are all well-defined. We shall also assume that $C_{\mathbf{y}}$ is nonsingular.

We remind the reader of the well-known “orthogonality principle” in standard Hilbert space theory. The orthogonality principle implies that the estimation error of the optimal estimator $\widehat{\mathbf{x}} - \mathbf{x}$ is uncorrelated with any function $g(\mathbf{y})$ of finite second moment. For a good reference, we recommend for instance the treatment by Luenberger [80, Chapter 3,4].

We first note that $\mathbb{E}[\widehat{\mathbf{x}}] = \overline{\mathbf{x}}$ since the desired estimator must be unbiased (take $g = 1$). We immediately get

$$b = \overline{\mathbf{x}} - W\overline{\mathbf{y}}. \tag{3.2}$$

Thus it remains to determine W. Taking $g(\mathbf{y}) = \mathbf{y} - \overline{\mathbf{y}}$, we have

$$\begin{aligned}
&\mathbb{E}[(\widehat{\mathbf{x}} - \mathbf{x})(\mathbf{y} - \overline{\mathbf{y}})^T] = 0, \\
&\Rightarrow \mathbb{E}[(W(\mathbf{y} - \overline{\mathbf{y}}) - (\mathbf{x} - \overline{\mathbf{x}}))(\mathbf{y} - \overline{\mathbf{y}})^T] = 0, \\
&\Rightarrow WC_{\mathbf{y}} - C_{\mathbf{x},\mathbf{y}} = 0, \\
&\Rightarrow W = C_{\mathbf{x},\mathbf{y}} C_{\mathbf{y}}^{-1}.
\end{aligned}$$

We may now determine the error covariance matrix

$$\begin{aligned}
C_e &\triangleq \mathbb{E}[(\widehat{\mathbf{x}} - \mathbf{x})(\widehat{\mathbf{x}} - \mathbf{x})^T] \\
&= \mathbb{E}[(\widehat{\mathbf{x}} - \mathbf{x})(W(\mathbf{y} - \overline{\mathbf{y}}) - (\mathbf{x} - \overline{\mathbf{x}}))^T] \\
&= 0 - \mathbb{E}[(\widehat{\mathbf{x}} - \mathbf{x})(\mathbf{x} - \overline{\mathbf{x}})^T] \\
&= \mathbb{E}[((\mathbf{x} - \overline{\mathbf{x}}) - W(\mathbf{y} - \overline{\mathbf{y}}))(\mathbf{x} - \overline{\mathbf{x}})^T] \\
&= C_{\mathbf{x}} - WC_{\mathbf{y},\mathbf{x}} \\
&= C_{\mathbf{x}} - C_{\mathbf{x},\mathbf{y}} C_{\mathbf{y}}^{-1} C_{\mathbf{y},\mathbf{x}}, \qquad (3.3)
\end{aligned}$$

where on the third line we used the orthogonality principle.

Let us now specialize to the situation here, where we have a linear observation model $\mathbf{y} = A\mathbf{x} + \mathbf{z}$, $\mathbf{x}, \mathbf{z}$ are uncorrelated, and A is a fixed matrix. Then, we have

$$\begin{aligned}
C_{\mathbf{y},\mathbf{x}} &= \mathbb{E}[(\mathbf{y} - \overline{\mathbf{y}})(\mathbf{x} - \overline{\mathbf{x}})^T] \\
&= \mathbb{E}[(A(\mathbf{x} - \overline{\mathbf{x}}) + (\mathbf{z} - \overline{\mathbf{z}}))(\mathbf{x} - \overline{\mathbf{x}})^T] \\
&= A\,\mathbb{E}[(\mathbf{x} - \overline{\mathbf{x}})(\mathbf{x} - \overline{\mathbf{x}})^T] + \mathbb{E}[(\mathbf{z} - \overline{\mathbf{z}})(\mathbf{x} - \overline{\mathbf{x}})^T] \\
&= AC_{\mathbf{x}}. \qquad (3.4)
\end{aligned}$$

Similarly,

$$\begin{aligned}
C_{\mathbf{y}} &= \mathbb{E}[(\mathbf{y}-\overline{\mathbf{y}})(\mathbf{y}-\overline{\mathbf{y}})^T] \\
&= \mathbb{E}[(A(\mathbf{x}-\overline{\mathbf{x}})+(\mathbf{z}-\overline{\mathbf{z}}))(A(\mathbf{x}-\overline{\mathbf{x}})+(\mathbf{z}-\overline{\mathbf{z}}))^T] \\
&= A\,\mathbb{E}[(\mathbf{x}-\overline{\mathbf{x}})(\mathbf{x}-\overline{\mathbf{x}})^T]A^T + \mathbb{E}[(\mathbf{z}-\overline{\mathbf{z}})(\mathbf{z}-\overline{\mathbf{z}})^T] \\
&= AC_{\mathbf{x}}A^T + C_{\mathbf{z}}. \qquad (3.5)
\end{aligned}$$

Substituting eqs. (3.4) and (3.5) into (3.3), we have the formula

$$C_e = C_{\mathbf{x}} - C_{\mathbf{x}}A^T(AC_{\mathbf{x}}A^T + C_{\mathbf{z}})^{-1}AC_{\mathbf{x}}. \qquad (3.6)$$

Now the expression for the LMMSE is simply $\mathrm{tr}(C_e)$. In our context here, we may use our assumption regarding the simultaneous diagonalization of all relevant matrices by the Fourier basis to simplify (3.6) in terms of spectral content as follows. Let us for convenience denote $\gamma \triangleq t/(n(W+\rho(\vec{a})J))$. We take (3.6) and rewrite it in terms of eigenvalues $\lambda(i)$ as

$$\sum_{i=0}^{n-1} \frac{\lambda_{C_{\mathbf{x}}}(i)\lambda_{C_{\mathbf{z}}}(i)}{\lambda_{C_{\mathbf{x}}}(i)\lambda_{AA^T}(i)+\lambda_{C_{\mathbf{z}}}(i)}. \qquad (3.7)$$

At this stage, we recall our assumptions and notation to see that $\lambda_{C_{\mathbf{x}}}(i) = \vec{d_i}$ (the prior on the scene), $\lambda_{AA^T}(i) = t^2|\widehat{\vec{a}}_i|^2/n^2$ (observation period of t, and the formula for eigenvalues of AA^T in terms of those for A), and $\lambda_{C_{\mathbf{z}}}(i) = t^2/(n^2\gamma)$ by our definition of the notational convenience γ and the formulation of the noise model.

Plugging in the above into (3.7), we finally obtain

$$\begin{aligned}
&\sum_{i=0}^{n-1} \frac{\vec{d_i}\frac{t^2}{n^2\gamma}}{\vec{d_i}\frac{t^2|\widehat{\vec{a}}_i|^2}{n^2} + \frac{t^2}{n^2\gamma}} \\
=&\sum_{i=0}^{n-1} \frac{1}{\frac{1}{\vec{d_i}} + \frac{t|\widehat{\vec{a}}_i|^2}{n(W+J\rho(\vec{a}))}},
\end{aligned}$$

which is nothing but (3.1) as claimed.

3.2.2 Model clarifications and intuition

In general, we assume $\vec{d_i} = (1/n)d(i/n)$ are n equally spaced samples from a nonnegative, bounded, continuous function $d(x)$ on $[0,1]$ with symmetry $d(x) = d(1-x)$ and normalized so that $d(0) = \theta$. For example, i.i.d. scenes correspond to $d(x) = \theta$. We note that our main result, Prop. 10, holds in greater generality. The above restriction on the form of $\vec{d}$ simply ensures correct physical scaling (invariant with respect to n) of the variance of total scene intensity coming from an arbitrary direction.

It is instructive to compare an ideal lens to a mask with respect to (3.1), as a function of exposure time. An ideal lens satisfies $\mathbf{A} = \mathbf{I}$, (i.e., $\vec{a} = (n, 0, \ldots, 0)$). Thus $\widehat{\vec{a}} = (n, n, \ldots, n)$. Then from (3.1), it can be readily seen that for a t growing arbitrarily slowly with n (say $t = \log(n)$), the LMMSE decays to 0 as $n \to \infty$. On the other hand, the entry-wise restriction $\vec{a} \in [0,1]$ that holds for a mask results in a significant reduction in $\|\widehat{\vec{a}}\|_2$. Due to this, in order to get an LMMSE that is bounded away from the trivial $\int d(x)\,\mathrm{d}x$ (corresponding to $t = 0, n \to \infty$), one needs an exposure time that is $\Omega(n)$. Of course, this is not surprising; there are strong benefits to lenses when they are available. The need for long exposure times for coded apertures is also a known phenomenon, consistent with the emphasis of [37] on “hypothesis tests” between scenes as opposed to resolving full detail.

One way to interpret increased t is that it reduces noise relative to the signal. All our main results established in the sequel (eqs. (3.9), (3.10) and (3.13)) show that one can construct apertures that are guaranteed to be tight within a constant factor of t. Under the above interpretation, what we establish rigorously is that our results are tight to within a constant number (≈ 18.30) of dB, regardless of the scene correlation structure given by $\vec{d}$. This factor may be read off from $2M(n)^2$ of Prop. 10.

3.3 Results

The goal of optimal aperture design (mathematically, optimal $\vec{a}$) is to minimize the LMMSE formula subject to the scene model, denoted as

$$m^*(n, t, W, J, \vec{d}) \triangleq \min_{\vec{a}} m(n, t, W, J, \vec{d}, \vec{a}).$$

Let us first understand why the minimization above is a challenging problem. Consider the even simpler problem in which the optimal transmissivity, say ρ_0, is given to us. Then, although $\vec{a} \in [0,1], \rho(\vec{a}) = \rho_0$ is a convex constraint, the LMMSE (3.1) which we wish to minimize is neither convex nor quasiconvex in $\vec{a}$, since $1/(1+cx^2)$ as a function of x lacks any of these behaviors.

In order to solve this problem, our general approach is as follows. First, we use Parseval's identity that relates time and frequency space. Under a fixed power budget, it is easy to solve for the optimal spectrum allocation $|\widehat{\vec{a}}_i|^2$ by studying the well-behaved and convex (as a function of x) $1/(1+cx)$ that has a solution given by waterfilling (3.8). Next, we are faced with the "coefficient problem" of finding a $\vec{a} \in [0,1]$ with given spectrum allocation. To address this, we appropriately apply a theorem of Nazarov [86, p. 5]. An exposition of Nazarov's work together with the context he draws from (e.g., the geometric ideas of [16], along with the analytic ideas of [33]) may be found in [15].

3.3.1 Lower bound

For notational ease, we let $\gamma \triangleq t/(n(W + J\rho))$ throughout.

We first derive a lower bound for LMMSE (3.1) based on waterfilling (see, e.g., [92, Thm 19.7]).

Lemma 21. *Consider the convex program*

$$\inf_{\sum P_i \le P, P_i \ge 0} \sum_{i=1}^{n-1} \frac{1}{\frac{1}{\vec{d}_i} + \gamma P_i}.$$

Then the solution is given by a "water-level" T implicitly governed by setting the optimal $P_i = \left(T - \frac{1}{\vec{d_i}}\right)^+ / \gamma$, satisfying $\sum_{i=1}^{n-1} P_i = P$.

Proof of Lemma 21. Let us associate Lagrange multipliers μ_i for the constraints $P_i \geq 0$, and λ for the constraint $\sum P_i \leq P$. Then we wish to minimize the Lagrangian

$$\lambda \left(\left(\sum_{i=1}^{n-1} P_i \right) - P \right) + \sum_{i=1}^{n-1} \left(\frac{1}{\frac{1}{\vec{d_i}} + \gamma P_i} - \mu_i P_i \right).$$

First order conditions yield

$$\lambda - \mu_i = \frac{\gamma}{\left(\frac{1}{\vec{d_i}} + \gamma P_i\right)^2}, \quad \mu_i P_i = 0.$$

Thus, for all i where P_i is not zero, $1/\vec{d_i} + \gamma P_i = T$ for some constant "water-level" T, or in other words, for such i,

$$P_i = \frac{T - \frac{1}{\vec{d_i}}}{\gamma}.$$

λ can't be zero for a nonzero P, so by complementary slackness we see that $\sum P_i = P$. This completes the proof. □

Remark 5. *The above convex program structure and associated "waterfilling" solution occur in a wide variety of problems and is well known. For example, the reference we provided [92, Thm 19.7] studies a similar problem to determine the capacity of a parallel AWGN channel. In the AWGN context, [93] attributes the derivation to [64]. Mathematically, all that we did was replace the convex $-\log(1 + ax)$ occuring in the capacity problem by the convex $1/(1 + ax)$ occuring here. We also note that a similar problem arises in the context of bit allocation for quantizers (see, e.g., [52, 8.2]), who attribute the solution (without the nonnegativity constraint) to [65].*

At a high level, what Lemma 21 allows us to do is solve for the best possible spectral allocation for coded apertures, assuming that one were able to set the levels freely modulo Parseval's equality relating time and frequency.

What we do next is observe that the lack of redirection of light for an aperture (essentially a l_∞ constraint on $\vec{a}$) allows one to write a natural upper bound on the l_2 norm of $\vec{a}$. The assumption on the lack of redirection of light turns into an upper bound on the l_2 norm of $\widehat{\vec{a}}$, and in turn allows us to apply Lemma 21. We give an elementary and short treatment below that is sufficient for our purposes as will become clear shortly. For the reader interested in the general conceptual framework, we recommend the study of the Fourier operator norm in the $l_p \to l_q$ sense for which there is a vast amount of material. One way to approach such a study is through tracing the historical line through the classical Hausdorff-Young inequality [127], [61], and the sharper Babenko-Beckner inequality [12], [17].

We accordingly have

Lemma 22. *Let $x_i \in [0,1], 1 \leq i \leq n$. Suppose $\sum x_i = r$, where $0 \leq r \leq n$. Then, $\sum x_i^2 \leq \lfloor r \rfloor + (r - \lfloor r \rfloor)^2$.*

Proof of Lemma 22. Think of a "trading mass" operation that takes a pair of distinct indices $(i,j), i \neq j$, with $x_i \leq x_j$ without loss of generality, and replaces x_i by $x_i - \epsilon$, x_j by $x_j + \epsilon$ where $\epsilon \geq 0$. This trading mass operation preserves the sum at r, but if $\epsilon > 0$, the sum of squares is strictly increased.

$$\begin{aligned}(x_i - \epsilon)^2 + (x_j + \epsilon)^2 &= (x_i^2 + x_j^2) + 2\epsilon(\epsilon + x_j - x_i) \\ &> x_i^2 + x_j^2.\end{aligned}$$

In such an operation, we may chose ϵ as large as possible until one of the x_i, x_j escapes $[0,1]$, and we term that as the "full mass trade" acting on (i,j).

Clearly any x maximizing the sum of squares must be invariant under the "full mass trade" across all pairs $(i,j), i \neq j$. In particular, at optimality, we can have at most one "intermediate" $x_i \in (0,1)$. These properties characterize the possible optimal vectors sufficiently to complete the proof.

For a less ad-hoc and more systematic proof, consider the set of vectors formed by permuting the entries of $(r - \lfloor r \rfloor, 1, 1, \ldots, 1, 0, 0, \ldots, 0)$ arbitrarily. These are the extremal vectors of the convex hull formed by these vectors, which in fact encompasses the entire constraints space. For a convex maximization problem, it is sufficient to examine the extremal vectors, and so we are once again done. One can view the first proof as an "algorithmic" variant of the second. □

Lemma 22 gives us the "power upper bound" that we need in order to employ waterfilling (Lemma 21) and prove

Proposition 7. *Let $\vec{a}$ satisfy $\rho(\vec{a}) = \rho$. Then*

$$m(n, t, W, J, \vec{d}, \vec{a}) \geq \frac{1}{\frac{n}{\theta} + \gamma n^2 \rho^2} + \sum_{i=1}^{n-1} \frac{1}{\frac{1}{\vec{d}_i} + \gamma P_i}. \tag{3.8}$$

Here $P_i = (1/\gamma)(T - 1/d_i)^+$ and total power $P = \sum_{i=1}^{n-1} P_i = n(\lfloor n\rho \rfloor + (n\rho - \lfloor n\rho \rfloor)^2) - n^2\rho^2$. Also note $P \leq n^2\rho(1-\rho)$. We remark that (3.8) *is sharp if and only if $|\widehat{\vec{a}}_i|^2 = P_i$ for $0 < i < n$.*

Proof. Recall that $\vec{d}_0 = \theta/n$, and that $\widehat{\vec{a}}_0 = n\rho$. This takes care of the first term. We now turn to the nonzero frequencies. First, note that $\sum_i \vec{a}_i^2 \leq n(\lfloor n\rho \rfloor + (n\rho - \lfloor n\rho \rfloor)^2)$ by Lemma 22. By Parseval's identity, we immediately have $\sum_i \widehat{\vec{a}}_i^2 \leq n(\lfloor n\rho \rfloor + (n\rho - \lfloor n\rho \rfloor)^2)$. We may substitute in $\widehat{\vec{a}}_0 = n\rho$ to get the expression for P above. The waterfilling Lemma 21 then gives the final expression.

The bound $P \leq n^2\rho(1-\rho)$ is convenient as it allows us to get rid of the floors for analytical study. We may readily derive it by

$$n(\lfloor n\rho \rfloor + (n\rho - \lfloor n\rho \rfloor)^2) - n^2\rho^2 \leq n(\lfloor n\rho \rfloor + n\rho - \lfloor n\rho \rfloor) - n^2\rho^2 \leq n^2\rho(1-\rho).$$

We used $0 \leq x - \lfloor x \rfloor \leq 1$ together with $x^2 \leq x$ on $[0, 1]$ above. □

Note that minimizing the right hand side over ρ gives a *lower bound* on $m^*(n, t, W, J, \vec{d})$. The minimization task is trivial numerically, but in general difficult analytically. We denote this optimal ρ by ρ^* henceforth.

3.3.2 Upper bound

Our goal here has been set from (3.8). Conceptually, the design issue is finding a $\vec{a} \in [0,1]$ with prescribed lower bounds $|\widehat{\vec{a}}_i|^2 \geq P_i$. In general, it is impossible to find such a $\vec{a}$ given arbitrary P_i satisfying the power bound of Proposition 7. Therefore, our lower bound (3.8) is not sharp in all settings. However, it should be noted that sharp cases do exist. Perhaps the conceptually simplest example is the analog of (3.8) for a lens, where our bound is sharp.

Our general approach is to back off by a factor C and obtain a $\vec{a} \in [0,1]$ with $|\widehat{\vec{a}}_i|^2 \geq P_i/C$. What we do next is address how we can guarantee such a C universally across $n, \vec{d}$. We shall move from simpler to more complex situations, and accordingly start off with i.i.d. scenes, where for infinitely many n one does not need the full generality of Nazarov's solution.

i.i.d. scenes

As clarified in our model, i.i.d. scenes correspond to $d(x) = \theta$, a constant. Examining the lower bound (3.8), we see that the waterfilling solution prescribes a $0, 1$ sequence with uniform spectrum allocation after the DC term ("spectrally flat sequences"). As already noted in [44, 125], one can certainly construct such spectrally flat sequences for infinitely many values of n, as long as they are at least asymptotically "unbiased" with $\rho = 1/2 - o(1)$. The performance of the apertures corresponding to these spectrally flat sequences meets the lower bound (as $n \to \infty$) as long as the optimal ρ^* is $1/2 - o(1)$ for the given t, W, J.

We now describe how one constructs spectrally flat sequences with $\rho = 1/2 - o(1)$. Our treatment here is brief and non-comprehensive. The construction we present here is based on elementary number theory, and can be attributed to Gauß [51].

First, let us reframe the problem of spectral flatness of a $\{0,1\}$ sequence of length n that we denote $\vec{a}$. Observe that $|\widehat{\vec{a}}_i|^2$ occurs as the Fourier transform of the (cyclic) autocorrelation sequence $b_j \triangleq \sum_{i=0}^{n-1} a_i a_{i-j}$, where we wrap indices modulo n. It thus clearly suffices to ensure that the autocorrelation sequence is constant for $1 \leq j \leq$

$n-1$. One can rephrase this as asking for a subset of $\mathcal{S} \subseteq \mathbb{Z}/n\mathbb{Z}$ (corresponding to the ones) with the property that the nonzero pair differences $i-j$ corresponding to pairs $(i,j) \in \mathcal{S} \times \mathcal{S}$ occur equally among $1, 2, \ldots, n-1$. For reasons that are intuitively clear from the physical situation, and will become mathematically clear shortly, we also want a "nontrivial" transmissivity, i.e. $\rho \in (0,1)$ as n grows. Indeed, the construction of Gauß is "unbiased" and achieves $\rho = 1/2 - o(1)$. We will first describe Gauß's construction below, and then show how one can generalize it and obtain $\rho \in \{1/4, 1/8\}$.

Let $n = p$ be a prime. Let us call $a \in \mathbb{Z}/n\mathbb{Z}$ a *quadratic residue* if $x^2 = a$ has a solution in $\mathbb{Z}/n\mathbb{Z}$, otherwise we call a a *quadratic nonresidue.* To make sure we are on the same page, $0, 1$ are always quadratic residues.

We assume the reader is familiar with Euler's criterion.

Lemma 23 (Euler). *Let $a \neq 0 \in \mathbb{Z}/p\mathbb{Z}$, where p is an odd prime. Then, in $\mathbb{Z}/p\mathbb{Z}$,*

$$a^{\frac{p-1}{2}} = \begin{cases} 1 & \text{if } a \text{ is a quadratic residue} \\ -1 & \text{otherwise.} \end{cases}$$

Proof. A proof may be found in almost any elementary number theory book, such as the freely available [108, Propn. 4.2.1]. □

Using Euler's criterion (Lemma 23), we may now easily prove the spectral flatness of Gauß's construction [51, Art. 356].

Theorem 7 (Gauß). *Let n be a prime p of the form $4k+3$. Let $\vec{a}_i = 1$ if i is a quadratic residue and $i \neq 0$, 0 otherwise. Then $\vec{a}$ is spectrally flat, i.e. $|\widehat{\vec{a}}_i|$ is constant for $1 \leq i \leq n-1$.*

Proof of Theorem 7. Taking $a = -1$ in Euler's criterion, we see that -1 is a quadratic nonresidue. Observe also that Euler's criterion immediately implies that the product of two quadratic nonresidues is a quadratic residue, a nonresidue and a residue is a nonresidue, and finally a residue and a residue is a residue.

Recall by our general discussion regarding autocorrelations and spectral flatness that our task amounts to showing that $r_a - r_b = \lambda$ has the same number of solutions in r_a, r_b across all $\lambda \neq 0 \in \mathbb{Z}/n\mathbb{Z}$, where $r_a, r_b \neq 0$ are quadratic residues.

Suppose λ is a quadratic residue. Note that $r_a - r_b = \lambda \leftrightarrow r_a r_b^{-1} - \lambda r_b^{-1} = 1$, where inverses are taken in the multiplicative group $(\mathbb{Z}/n\mathbb{Z})^*$. This observation gives a bijection between difference pairs for λ and pairs for 1. Thus the number of solutions is the same across all λ that are quadratic residues.

Now, suppose λ is a quadratic nonresidue. Then $-\lambda$ is a quadratic residue, and we have $r_a - r_b = \lambda \leftrightarrow r_b - r_a = -\lambda$. Thus the number of solutions for quadratric residues is the same as those for nonresidues when we combine with the preceding paragraph. This completes the proof. □

Remark 6. *The astute reader may have observed that we did not actually compute the DFT of Gauß's quadratic residue sequence, as we did not need to determine the phase for checking spectral flatness. By comparsion, Gauß [51, Art. 356] in fact computed the DFT of such a sequence (with the phase information) explicitly in both the cases $p = 4k+1$ and $p = 4k+3$. A popular approach for performing the full DFT computation due to Dirichlet is the Poisson summation formula, see, e.g., [32, Ch. 2].*

Now that we have established the existence of "unbiased" spectrally flat sequences (at least for some n), a natural question is how good is using an "unbiased" spectrally flat sequence when $\rho^* \neq 1/2$? The answer is given in the following

Proposition 8. *Let θ, W, J be fixed and let $\vec{d} = (\theta/n)\vec{1}$. Then for infinitely many n, there exists a $\vec{a} \in [0,1]$ such that*

$$m(n, 2t, W, J, \vec{d}, \vec{a}) \leq m^*(n, t, W, J, \vec{d}). \tag{3.9}$$

In other words, "unbiased" spectrally flat sequences are always guaranteed to achieve optimal LMMSE at the expense of increasing the exposure time t by a factor of at most 2. In the sequel, we show how one can reduce this factor even further.

The improvement of the constant factor 2 is achieved by using spectrally flat sequences with $\rho = 1/8 - o(1)$ and $\rho = 1/4 - o(1)$ in combination with $\rho = 1/2 - o(1)$, and allows us to refine 2 to 8/7. Intuitively, a lower ρ helps with scenarios where shot noise dominates over thermal noise, and a higher ρ when thermal noise dominates shot noise. One therefore expects that having multiple constructions with different ρ helps with aperture selection, tailored to the specific shot noise versus thermal noise scenario at hand. If one had other values of ρ with spectrally flat constructions, one could possibly reduce the constant factor further, depending on the calculation presented in the upcoming Lemma 24.

The construction of spectrally flat sequences for $\rho \in \{1/4, 1/8\}$ is based on well-established cyclotomic number computations originating in [51] and developed further in [38] in number theory. $\rho = 1/4$ corresponds to quartic residues [24], and $\rho = 1/8$ corresponds to octic residues [74]. It should be emphasized, however, that in contrast to the case in which $\rho = 1/2$, the existence of such sequences for infinitely many values of n is not guaranteed, because no single-variable quadratic taking on infinitely many primes is known [117].

Of perhaps greater importance is the fact that the octic residue constructions of [74] rely upon primes that come from a second order linear recurrence with rather large coefficients, arising as the solutions of Brahmagupta-Pell equations. There is thus a paucity of such constructions; indeed [74] gives only two such n below 10^9, namely $n = 73$ and $n = 26041$. On the other hand, the quartic residue constructions are reasonably numerous, with over 150 of them available below 10^7. Even restricting ourselves to the quartic residues allows us to tighten from 2 to 4/3. Summarizing all of the above, we have

Proposition 9. *Let θ, W, J be fixed and let $\vec{d} = (\theta/n)\vec{1}$. Then for some values of n that exist even beyond, e.g., 10^9, there exists a $\vec{a} \in [0,1]$ such that:*

$$m(n, (8/7)t, W, J, \vec{d}, \vec{a}) \leq m^*(n, t, W, J, \vec{d}). \tag{3.10a}$$

Moreover, for many (> 150 for $n < 10^7$) values of n that exist even beyond, e.g., 10^9, *there exists a $\vec{a} \in [0,1]$ such that*

$$m(n,(4/3)t,W,J,\vec{d},\vec{a}) \leq m^*(n,t,W,J,\vec{d}). \tag{3.10b}$$

Before turning to the proof, we have a few words to say about the constants $2, 4/3, 8/7$. The astute reader may have noticed that they were obtained for $\rho \in \{1/2, 1/4, 1/8\}$, and may have also noticed the $2^k/(2^k-1)$ pattern. One may thus hope for this pattern to continue if one could (hypothetically) construct spectrally flat sequences with $\rho = 1/16$ as well. Unfortunately, this is not the case. In fact, the values of these "magic" constants come from a two variable optimization that is readily carried out on a computer, but is somewhat painful and certainly unilluminating to do by hand. We obtain the constants in

Lemma 24. *Let $a > 0$, and let*

$$f_a(\rho) : [0,1] \to \mathbb{R} \triangleq \frac{\rho(1-\rho)}{a+\rho}.$$

Let

$$M(a,\mathcal{B}) \triangleq \max_{x\in[0,1]} \frac{f_a(x)}{\max_{\rho\in\mathcal{B}} f_a(\rho)}.$$

Then

$$\begin{aligned} M\left(a, \left\{\frac{1}{2}\right\}\right) &\leq 2, \\ M\left(a, \left\{\frac{1}{2}, \frac{1}{4}\right\}\right) &\leq \frac{4}{3}, \\ M\left(a, \left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}\right\}\right) &\leq \frac{8}{7}. \end{aligned}$$

Proof of Lemma 24. See Appendix A. □

We also note that the constants are rather small, especially compared with those arising from the general solution based on Nazarov's theorem as will become clear later. A high level takeaway from Lemma 24 is that having different aperture designs with different transmissivities allows one greater design flexibility, and thus allows one to tailor the choice to the relative shot/thermal noise levels. The precise numerical factors are much less important.

One may wonder what the optimal (in the min-max sense) choice of transmissivity is within our model. After all, the mathematics of minimizing the multiplicative factor subject to a single choice of ρ is clear. Along the lines of the proof of Lemma 24, this is a straightforward calculus exercise. The answer turns out to be $\rho = 1/4$, yielding a factor of 4/3. In other words, in the notation of Lemma 24, we have for any $\rho \in [0, 1]$,

$$M\left(a, \left\{\frac{1}{4}\right\}\right) \leq \frac{4}{3} \leq \sup_a M\left(a, \{\rho\}\right). \tag{3.11}$$

We do not recommend that the reader pays too much attention to the above (3.11). After all, in general, we believe such questions are best left to actual physical tests as there are a number of factors our model ignores.

We now turn to the proof of Props. 8, 9.

Proof of Props. 8, 9. Recall that we wish to use spectrally flat sequences. First, we note that the indicator/characteristic function of the "difference sets" of [24, 74] are in our language spectrally flat sequences. The constant factor is given by the following single variable optimization. In view of (3.8), let $f_a(\rho) = (\rho(1-\rho))/(a+\rho)$ defined on $[0, 1]$; a corresponds to W/J. The numerator comes from the power bound, the denominator from the noise penalty. Then, $M(a, \rho) = \sup_x f_a(x)/f_a(\rho)$ is the multiplicative loss factor for a fixed W/J and fixed $\rho \in \{0.125, 0.25, 0.5\}$. One may then optimize over ρ, a to get the constant (3.10a) via Lemma 24. This proof, modified to $\rho \in \{0.25, 0.5\}$ and $\rho \in \{0.5\}$, also yields (3.10b) and (3.9) respectively, again by the computations of Lemma 24. The fact that there are infinitely many n for $\rho = 0.5 - o(1)$ follows from the quadratic residue construction together with the well known fact that there are infinitely many primes $p = 4k+3$ (see, e.g., [10, Ch. 7]). □

Correlated scenes

We now turn to correlated scenes. Here the waterfilling is nontrivial, and prescribes an unequal spectrum allocation. We therefore invoke Nazarov's solution to the coefficient problem [86, p. 5], and also provide a statement here specialized to the DFT and l_∞ that we use.

Nazarov's theorem [86, p. 5] is presented in a very elegant and concise manner. We do not know of any nontrivial improvements to it, either in exposition or in power. The proof is sufficiently short that we present it here. We caution the reader that although short, this proof is certainly nontrivial and can be a bit mysterious. In fact, a lot of Nazarov's original paper [86] is devoted to "unravelling" the mysterious steps, as alluded to in the epigraph.

Consider the problem stated in the epigraph, which is Tarski's famous plank problem [2]. A strip is a region enclosed between two parallel hyperplanes. The width of a convex body is defined as the width of the narrowest strip containing it. Tarski asked whether given a convex set $B \in \mathbb{R}^n$, is it possible to cover it by several strips of total width less than the width of B? The answer is no, but it was surprisingly difficult to prove. Nevertheless, Bang's solution [16] is very short (2 pages) and completely elementary! Nazarov includes Bang's solution in his paper [86] for the reader's convenience.

Let us think about the nature of the plank problem and why one might see some links to the coefficient problem. A strip centered at the origin may be written as $\{x : |\langle x, \psi \rangle| \leq a\}$, where ψ is a unit vector. Thus a point y in some convex set B that isn't covered by a collection of strips with unit normals $\psi_1, \ldots, \psi_n$ and coefficients $a_1, \ldots, a_n$ must have large coefficients with respect to all these vectors simultaneously. As such, it is not unreasonable to expect that the methods of Bang [16] that understood (and in fact explicitly constructed) points y that are not covered by the strips could potentially be used to construct vectors that have large coefficients with respect to an orthonormal basis. The problem of constructing such vectors with large coefficients ties in well with the precise design problem we face here based on the belief that waterfilling should guide the spectral allocation.

It is one thing to spot the above heuristic link between covering a set by strips and

[2] The problem apparently first appeared in print in 1932. See Bang's original paper [16] for a reference and historical remarks.

the coefficient problem, a nontrivial ask in of itself, and another to actually solidify the link. Nazarov succeeded in [86], and we present a fruit of his labor in

Theorem 8 (Nazarov)**.** *Let T be a measure space with probability measure μ. Let $\psi_j : T \to \mathbb{R}$ be an at most countable system of functions satisfying*

$$\left|\sum_j c_j \psi_j\right|_2 \leq \left(\sum_j c_j^2\right)^{\frac{1}{2}},$$

for any $c_j \in \mathbb{R}$. Suppose $2 \leq p \leq \infty$, and let q be the conjugate exponent to p, i.e., $p^{-1} + q^{-1} = 1$. Assume

$$\forall j, \quad |\psi_j|_q \geq \beta > 0.$$

Let $0 \leq p_i$ satisfy $\sum_i p_i = 1$. Then there exists $b \in l_p(T)$ with

$$|b|_p \leq \left(\frac{3\pi}{2}\right)^{1-\frac{2}{p}} \beta^{-2},$$

such that

$$\forall j, \quad |\langle b, \psi_j \rangle| \geq \sqrt{p_j}.$$

Here, inner products are defined with respect to the probability measure μ by

$$\langle f, g \rangle \triangleq \int_T f g d\mu.$$

Proof of Theorem 8. Let $\boldsymbol{\epsilon} \in (\pm 1, \pm 1, \dots)$ [3], and define

$$f_{\boldsymbol{\epsilon}} \triangleq \sum_j \boldsymbol{\epsilon}_j \sqrt{p_j} \psi_j.$$

Define $\Phi(x)$ via

$$\Phi''(x) = (1 + x^2)^{\frac{2}{p}-1}, \Phi(0) = \Phi'(0) = 0.$$

Such a function exists and is uniquely determined by the existence and uniqueness

[3] Nazarov calls $\boldsymbol{\epsilon}$ a “sign cortège”.

theorem for differential equations.

Now the integral $I(f) \triangleq \int_T \Phi(f) d\mu$ is well defined and continuous in $l_2(T)$. Since the family $\{f_{\boldsymbol{\epsilon}}\}$ is compact in the topology of $l_2(T)$, one can find a cortège $\boldsymbol{\epsilon}^*$ such that $f_{\boldsymbol{\epsilon}^*}$ **maximizes** $I(f)$ over all $f_{\boldsymbol{\epsilon}}$.

Now consider the cortège obtained by "flipping" a single sign, and accordingly define "bit-flipped" $f_j = f_{\boldsymbol{\epsilon}^*} - 2\boldsymbol{\epsilon}_j^* \sqrt{p_j} \psi_j$. By the mean value theorem and our choice of $\boldsymbol{\epsilon}^*$, we have

$$\begin{aligned} 0 &\leq \int_T (\Phi(f_{\boldsymbol{\epsilon}^*}) - \Phi(f_j)) d\mu \\ &= \int_T \Phi'(f_{\boldsymbol{\epsilon}^*})(f_{\boldsymbol{\epsilon}^*} - f_j) d\mu + (1/2) \int_T \Phi''(g)(f_{\boldsymbol{\epsilon}^*} - f_j)^2 d\mu, \end{aligned}$$

where g lies between f, f_j pointwise. Recalling the definition of f, f_j, we obtain

$$\left| \int_T \Phi'(f_{\boldsymbol{\epsilon}^*}) \psi_j d\mu \right| \geq \sqrt{p_j} \int_T \Phi''(g) \psi_j^2 d\mu. \tag{3.12}$$

At this stage, the path forward is as follows. We will use a (possibly scaled) version of $\Phi'(f_{\boldsymbol{\epsilon}^*})$ as the desired b. In order to do so, we will accomplish two tasks:

1. Give a uniform lower bound on $\int_T \Phi''(g) \psi_j^2 d\mu$. We claim that $\beta^2 3^{\frac{2}{p}-1}$ works.

2. Show that $\Phi'(f_{\boldsymbol{\epsilon}^*}) \in l_p(T)$. We claim that $|\Phi'(f_{\boldsymbol{\epsilon}^*})|_p \leq \left(\frac{\pi}{2}\right)^{1-\frac{2}{p}}$.

Let us look at the first task. Here, recall that $\Phi''(x) = (1+x^2)^{\frac{2}{p}-1}$, so by Hölder's inequality [4], we have

$$\left(\int_T \Phi''(g) \psi_j^2 d\mu \right)^{\frac{q}{2}} \left(\int_T (1+g^2) d\mu \right)^{1-\frac{q}{2}} \geq \left(\int_T |\psi_j|^q d\mu \right).$$

But $\int_T (1+g^2) d\mu \leq \int_T (1 + f_{\boldsymbol{\epsilon}^*}^2 + f_j^2) d\mu = 3$, so $\int_T \Phi''(g) \psi_j^2 d\mu \geq \beta^2 3^{\frac{2}{p}-1}$.

[4] Historical note: Although commonly called Hölder's inequality, Rogers discovered it in 1888 [99] before Hölder in 1889 [63]. Henceforth we follow the common convention.

Now for the second task. We give an upper bound on $|\Phi'(x)|$ by Hölder's inequality.

$$|\Phi'(x)| = \int_0^{|x|} (1+s^2)^{\frac{2}{p}-1} ds \leq \left(\int_0^{|x|} ds \right)^{\frac{2}{p}} \left(\int_0^{|x|} (1+s^2)^{-1} ds \right)^{1-\frac{2}{p}}$$
$$\leq \left(\frac{\pi}{2}\right)^{1-\frac{2}{p}} |x|^{\frac{2}{p}}.$$

Thus

$$|\Phi'(f_\epsilon)|_p \leq \left(\frac{\pi}{2}\right)^{1-\frac{2}{p}},$$

completing the second task.

Take $b = 3^{1-\frac{2}{p}} \beta^{-2} \Phi'(f_{\epsilon^*})$ to complete the proof. □

Remark 7. *We note that the constant $3\pi/2$ for $p = \infty$ is not sharp and may be improved. A cheap way of doing this is studying the family of functions $\Phi_b(x)$ governed by $\Phi_b''(x) = 1/(1+bx^2)$ and repeating Nazarov's argument. It turns out that $b = 1/2$ optimizes the constant in this family, and yields the very modest improvement of $3\pi/2 \to \sqrt{2}\pi$.*

We return to our question of coded aperture design, the associated Fourier analysis on $\mathbb{Z}/n\mathbb{Z}$, and specialize Nazarov's theorem 8 to such a situation. First, let us define inner products with respect to the uniform probability distribution on $\{0, 1, \ldots, n-1\}$. Let $0 \leq i, j \leq n-1$, and let ψ_j be a orthonormal basis for the DFT on real sequences. Explicitly, let $h = \lceil (n-1)/2 \rceil, \omega = 2\pi/n$. Let $\psi_0(i) = 1$, $\psi_j(i) = \sqrt{2}\cos(\omega j i)$ for $0 < j < h$, $\psi_j(i) = \sqrt{2}\sin(\omega j i)$ for $h < j < n$. If n is even, let $\psi_h(i) = \cos(\omega h i)$, otherwise $\psi_h(i) = \sqrt{2}\cos(\omega h i)$. Finally, let $\beta(n) = \min_j |\psi_j|_1$.

Corollary 7 (Nazarov)**.** *Let $M(n) = ((3\pi)/2)\beta(n)^{-2}$. Let $0 \leq p_0, p_1, \ldots, p_{n-1}$ be such that $\sum p_j = 1$. Then there exists a $\vec{b} \in [-M(n), M(n)]$ with $|\langle \vec{b}, \psi_j \rangle|^2 \geq p_j$ for all $0 \leq j \leq n-1$.*

Proof of Corollary 7. Take $p = \infty$ and use ψ_i as an orthonormal basis in Nazarov's Theorem 8. □

With Corollary 7 in hand, we are able to reach a far more general version of Prop. 8, 9 valid for any n and any scene prior $\vec{d}$. Also, in Sec. 3.3.3 we show how to construct sequences that achieve our goal of being guaranteed to lie within a constant (independent of $n, \vec{d}$) factor of optimal sequences. At the moment, we know that their existence is guaranteed by Corollary 7.

Proposition 10. *For all $n, t, W, J, \vec{d}$, there exists a $\vec{a} \in [0, 1]$ such that*

$$m(n, 2M(n)^2 t, W, J, \vec{d}, \vec{a}) \leq m^*(n, t, W, J, \vec{d}). \tag{3.13}$$

Furthermore, we have

$$M(n) \in [(3\pi^3)/16 + o(1), 3\pi + o(1)]. \tag{3.14}$$

The justification of the tightness of (3.13) lies in establishing (3.14), which we do first. The phenomenon is captured by the factorization of n, with the best, that is the largest, β occurring for n prime, and the worst occuring for n divisible by 4.

In order to elucidate this behavior, we first establish the following

Lemma 25. *Let $n \geq 4$ be a natural number, and let $\omega \triangleq 2\pi/n$. Consider A_n, B_n defined via*

$$\begin{aligned} A_n &\triangleq \frac{1}{n}\sum_{k=0}^{n-1} |\cos(\omega k)| \\ B_n &\triangleq \frac{1}{n}\sum_{k=0}^{n-1} |\sin(\omega k)|. \end{aligned}$$

Then, for $n \geq 4$,

$$\begin{aligned} A_n, B_n &\geq \frac{1}{2}, \\ A_n, B_n &= \frac{2}{\pi} + O\left(\frac{1}{n}\right). \end{aligned}$$

Furthemore, $1/2$ is attained at $n = 4$ only.

Proof of Lemma 25. We have the well known triangle inequality written as

$$||x| - |y|| \leq |x - y|.$$

Thus, we have

$$||\sin(x)| - |\sin(y)|| \leq |\sin(x) - \sin(y)| \leq \left(\sup_x |\sin'(x)|\right) |x - y| = |x - y|.$$

Similarly, we get $||\cos(x)| - |\cos(y)|| \leq |x - y|$. As such, by classical results on the comparison of a (left) Riemann sum with the definite integral (see e.g. [100, Ch. 6]), we get

$$\left|A_n - \int_0^1 |\cos(2\pi x)| dx\right| \leq \frac{2\pi}{n},$$
$$\left|B_n - \int_0^1 |\sin(2\pi x)| dx\right| \leq \frac{2\pi}{n}.$$

The conceptual part of the proof is completed by the above discussion. The rest depends on the following exercise, that we recommend performing on a computer: compute A_n, B_n up to say $n = 10000$, and check the minimum, which happens to occur at $n = 4$. For $n > 10000$, A_n, B_n are certainly within say 0.01 of the asymptotic $2/\pi \approx 0.6366$ established above. □

Remark 8. *Our published work [8] had some remarks regarding Euler-Maclaurin summation in the proof sketch given there. The point we were emphasizing there is that Lemma 25 and similar such statements are entirely routine matters that fall under the general umbrella of Euler-Maclaurin summation, which provides an asymptotic expansion for the difference between a sum and an integral in terms of higher derivatives evaluated at the endpoints. However, the function $|\cos(x)|$ has annoying discontinuities in the first derivative. Nevertheless, such matters are also entirely routine and can be handled in numerous ways. Our work [8] gave an ad-hoc one via smoothing out the discontinuities with splines.*

We came up with the much clearer and simpler approach via the triangle inequal-

ity (or more generally the "modulus of continuity") after publication. Nevertheless, we still believe that Euler-Maclaurin summation is well worth understanding for the following reasons.

1. *The reader may justifiably view our use of the triangle inequality as a mere trick that allowed us to get what we need and may thus yearn for a conceptual framework. Euler-Maclaurin summation provides one route.*
2. *The reader may wish to obtain fine grained control on the error terms that we simply lumped into the $O(1/n)$ of Lemma 25. Euler-Maclaurin gives a full asymptotic expansion.*

We therefore recommend [56, Sec. 9.5] for an easy to read treatment suitable for a wide audience, and/or a blog post of Tao [111] for an exposition at a more sophisticated level together with applications to number theory.

With the routine Lemma 25 in hand, we have the following Lemma which establishes (3.14).

Lemma 26. *Let $\beta(n)$ denote the l_1 lower bound of Nazarov's theorem 8, specialized to the real orthonormal basis ψ_j for the DFT defined earlier. Then,*

$$\beta(n) \in \left[\frac{1}{\sqrt{2}} + o(1), \frac{2\sqrt{2}}{\pi} + o(1)\right]$$

as $n \to \infty$. Moreover, if we restrict to n being prime,

$$\beta(n) = \frac{2\sqrt{2}}{\pi} + o(1).$$

Proof of Lemma 26. Let us first examine the case in which $n = p$, where p is a prime. Then, for any $j \neq 0$, jk sweeps over $\{0, 1, \ldots, p-1\}$, modulo p as k sweeps over $\{0, 1 \ldots, p-1\}$. Thus, we are examining the Riemann sum approximation

$$\frac{1}{p}\sum_{k=0}^{p-1} \cos\left(\frac{2\pi k}{p}\right)$$

to $\int_0^1 |\cos(2\pi x)|dx = 2/\pi$. The l_2 norm of $\cos(2\pi x)$ on $[0,1]$ is $1/\sqrt{2}$, and we may invoke Lemma 25 to complete the proof.

The composite case is slightly more involved, as it needs to take into account the divisor structure of n, which prevents such symmetry of the cosine vectors. What we mean by this is the following. Consider the basic term jk/n, where j is fixed, and k varies over $0, 1, \ldots, n-1$. We may divide by the greatest common divisor (gcd) (j, n) to equivalently study $j'k/n'$, where j', n' are coprime ($(j', n') = 1$). Now as k varies over $0, 1, \ldots, n-1$, k varies over $0, 1, \ldots, n'-1$, modulo n', with each residue occuring the same number of times. As j', n' are coprime, $j'k$ also varies over $0, 1, \ldots, n'-1$, modulo n', with each residue occuring the same number of times. Thus, one is simply looking at a Riemann sum approximation to an integral number (n/n') of periods of $|\cos(2\pi x)|$. Thus we may once again invoke Lemma 25 to complete the proof: the role of n is replaced by that of its possible divisors. In particular, the "worst" possible divisor of 4 from Lemma 25 results in the "worst" β for n divisible by 4. □

We emphasize that by Lemma 26 $M(n) \leq C$ for some constant $C \approx 9.4248$, with even better values available at, e.g., prime $n > 100$. There, $C \approx 5.8146$ suffices.

Proof of Prop. 10. Corollary 7, with $p_0 = 0$ and $p_j = P_j / \sum_j P_j$ for $0 < j < n$ yields a $\vec{b}$ with $|\vec{b}|_\infty \leq M(n)$ and $|\langle \vec{b}, \psi_j \rangle|^2 \geq p_j$ for $0 < j < n$. Without loss, we may assume that $\langle \vec{b}, \psi_0 \rangle \leq 0$, else simply flip signs. The fact that we have the appropriate frequency magnitudes for ψ_j translates to an equivalent statement for the complex exponentials by considering the real and imaginary parts separately. Recalling the upper bound $P \leq n^2 \rho(1-\rho)$, we obtain $|\widehat{\vec{b}}_j|^2 \geq P_j/(\rho(1-\rho))$. Now consider $\vec{a} = (\vec{b} + M(n))/2M(n)$. Then, $\vec{a} \in [0,1]$, $\rho(\vec{a}) \leq 0.5$, and $|\widehat{\vec{a}}_j|^2 \geq P_j/(4M(n)^2\rho(1-\rho))$ for $0 < j < n$. We are now in a similar situation to that of Prop. 8, except with an extra $M(n)^2$ factor, and the fact that $\rho(\vec{a}) \leq 0.5$ instead of $\rho(\vec{a}) = 0.5 + o(1)$. The latter is no problem, as lower ρ only helps us with the shot noise term, and the former simply multiplies the 2 of (3.9) by $M(n)^2$. □

3.3.3 Greedy algorithm

Here we propose a (heuristically) efficient algorithm to construct vectors $\vec{a}$ that satisfy the conditions of Prop. 10. This algorithm has its roots in Nazarov's original proof. Recall that at a high level, Nazarov's theoretical construction boils down to finding a "sign cortège" [86, p. 6] that is globally optimal for a certain real-valued Boolean function of n signs, taking exponential time in the worst case. However, notice that in the proof of Theorem 8, we emphasized the word "**maximizes**". A closer examination of the proof of Theorem 8 reveals that one simply needs a sign cortège that is locally optimal in the sense of Hamming geometry for the proof to work! In less explicit terms, one can simply replace the emphasized "**maximizes**" with "**locally maximizes**" and the proof still goes through.

Our observation suggests a natural greedy algorithm where one starts with a random cortège, and then flips one sign at a time if it improves the objective, repeating until no further improvement is possible. We do not know of any theoretical justification as to why this is a good algorithm: for instance, a function on the hypercube may have only one local optimum.

Nevertheless, in our simulations [5] this runs very fast, and empirically takes time cubic in n. For example, on our standard laptop, we can generate apertures for $n = 2000$ in 4 seconds. Our situation superficially resembles the situation of the simplex algorithm and the smoothed analysis of [105], or more directly recent work on max-cut [9]. Direct application of the methods of [9] to obtain theoretical guarantees runs into difficulties with the nonlinear change in objective with a single bit flip in our setting, unlike the linear change for max-cut. As such, we defer theoretical study of the greedy algorithm given here to future work.

3.3.4 Simulations

We give a simple illustration in Fig. 3-5 which confirms the following intuition based on our main results eqs. (3.9), (3.10) and (3.13). With an i.i.d. scene prior, one

[5] Code:`https://github.com/gajjanag/apertures`

would prefer using the spectrally flat construction as opposed to the one coming from Nazarov's theorem due to the smaller constant. On the other hand, with a strong prior—e.g., a bandlimited one—the waterfilling becomes highly skewed, and one would favor the one coming from Nazarov's theorem as it takes into account such strong skewing of the desired spectrum and accordingly utilizes the spectrum better. For completeness, we also include the performance of a random on-off sequence with density ρ [125], where ρ is optimized over $[0, 1]$ for each t.

In Fig. 3-5, we note that the Nazarov and lower bound plots are within a constant distance of each other even as t grows. The constant distance arises from our notion of "near-optimality", namely that the performance is within a constant multiplicative factor of being optimal. That multiplicative factor translates to a constant on a logarithmic scale. For the spectrally flat case, the plot does not diverge from the lower bound as t grows, though the gap between the two depends on the choice of prior, and can be arbitrarily large given certain priors unlike the Nazarov plot. We also note that the optimal random on-off (where ρ is optimized for each value of t) diverges from the lower bound. We may heuristically justify the phenomenon as follows. We may assume that asymptotically, $|\widehat{a}_i|^2$ behaves like a χ-squared random variable of two degrees of freedom, and plug in its density into the LMMSE expression (3.1). Upon performing the computation, this divergence ultimately comes from the fact that

$$\int_0^\infty \frac{ae^{-x}dx}{1+ax} = \Theta(\log(a)), \quad a \to \infty.$$

3.4 Discussion and Future Work

Our refined analysis of a model drawing heavily from [125] yields tight conclusions across all scene correlation patterns and noise regimes, with sharp conclusions available in some specific scenarios. Moreover, we give heuristically efficient algorithms for the generation of optimal coded apertures. We also note that similar conclusions to our main results eqs. (3.9), (3.10) and (3.13) also hold for MI and Gaussian statistics of [125], simply because of the form of the expression for MI. Basically, MI is another functional that may be written in the form

$$\sum_{i=0}^{n-1} f(|\widehat{a}_i|),$$

where f is concave.

Nazarov's theorem is sufficiently general to allow the analysis of such functionals to a similar extent to what we did for LMMSE. Namely, although we can't necessarily give sharp answers (in our view, a difficult problem!), we can give answers that are effectively constructible and guaranteed to be good in the sense of being a constant factor away from optimal. Naturally, the precise sense of this, and the exact constant factor, will depend on the choice of f. We leave such exercises to the interested reader, who may have his/her own favorite f.

Furthermore, we note that our conclusions generalize naturally to 2D apertures, and in particular we have a tight characterization of optimal coded apertures in that setting. Concretely, one simply needs to take $\beta(n)^2$ as opposed to $\beta(n)$ due to the squaring of the l_1 lower bound for the 2D DFT. The rest of the analysis of Theorem 8 and Prop. 10 carries over naturally, with the orthogonal basis provided by products $\psi_j \otimes \psi_k$. We emphasize that this works regardless of the scene prior, even ones which are not separable. With an i.i.d. prior, separable apertures are optimal up to constants as in 1D, and in fact taking a product of spectrally flat apertures yields natural analogs of Props. 8, 9. However, with other priors, it seems like one needs the generality provided by Theorem 8. For separable priors, one can simply use a product of apertures arising from the specialization of Nazarov's theorem to the 1D DFT that we described in this chapter. For general priors, one can repeat the analysis of this chapter, applied to the 2D DFT instead with basis $\psi_j \otimes \psi_k$. Our work thus also answers the question of 2D apertures raised in [125]. We also view experimental verification of these ideas as a worthwhile task.

As noted in [125], [119] raises the question of whether continuous-valued masks perform better than binary-valued ones. Our work sheds some light on this: the solution of Nazarov which we have shown is tight does seem to use the flexibility of the l_∞ norm in an essential way; see, e.g., [58, p. 12] for more on this. And more specifically, we have numerical evidence for finite n; to give a concrete example, for $n = 13$, the mask $[1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0]$ has optimal LMMSE for an i.i.d. scene over binary-valued masks for $\rho = 6/13, \theta = 0.01, W = J = 0.001, t = 130$, but is improved upon by the continuous-valued mask whose first entry is equal to ϵ and whose ith entry is equal to $1 - \epsilon/6$ if $i - 1$ is a quadratic residue modulo 13, and 0 otherwise, for $0.26 \leq \epsilon \leq 0.34$. We do view a full resolution of this question to be of significant mathematical as well as engineering interest. The engineering interest stems from our belief that partial occluders may be more difficult to synthesize than the classical on-off apertures. The surrounding mathematical landscape is rich, and a touchstone is perhaps provided by the recently resolved (by [13]) "Littlewood's flatness" problem. First raised by Erdős [41, Prob. 26] and later extended and popularized by Littlewood in several of his papers such as [77] as well as his book [78], Littlewood's problem is very simple to state. In the original form proposed by Erdős [41, Prob. 26], we have

Theorem 9. *There exists, for each n, a polynomial*

$$f_n(z) = \sum_{k=1}^{n} \epsilon_{n,k} z^k \quad (\epsilon_{n,k} = \pm 1),$$

such that, for all θ, $c_1\sqrt{n} < |f_n(e^{i\theta})| < c_2\sqrt{n}$, where c_1, c_2 are positive constants independent of θ, n.

Note that what we ask here, for the spectrally flat case, is a weaker variant of the above Theorem 9, where we restrict θ to $2\pi k/n$ with $k = 0, 1, \ldots, n-1$. However, analogous questions can be asked in the spirit of Nazarov's solution to the coefficient problem, where one still restricts θ to $2\pi k/n$, but asks for a "shaped" magnitude response across the unit circle. We also think that an efficient numerical procedure (at least heuristically), both in the context of Littlewood flatness as well as a "shaped" magnitude response in the spirit of Nazarov's solution is an interesting challenge.

We note that our weaker variant of the flatness problem has some aspects that may be understood without too much investment. For example, allowing the $\epsilon_{n,k}$ to lie on the unit circle allows for a trivial solution to the weak variant above for

$n = p$ prime via the Gauß sum. One can also work with the standard FFT recursion (Cooley/Tukey/Gauß) and construct solutions inductively to the weak variant for $n = 2^k$ with $\epsilon_{n,k} \in \{\pm 1, \pm i\}$. Naturally, the full resolution by [13] is much more involved and we do not describe it here.

We also note a potentially interesting approach towards understanding the limits to which the coefficient problem can be solved with binary vectors that draws a connection with our work in the previous chapter, and linear programming bounds (2.12) more specifically. Briefly, a $\{0,1\}$ vector of length 2^d can be thought of as a code in the Hamming space $\{0,1\}^d$. As the discussion in Remark 1 makes clear, the dual distance distribution of this code is nothing but the (binary) Fourier squared magnitude, grouped by Hamming weight. Thus, constraints on the distance/dual distance distribution translate into constraints on the (binary) Fourier coefficients. One may object that the characters are not the same as those for $\mathbb{Z}/n\mathbb{Z}$, however we emphasize that Nazarov's solution works for both and thus view this as a reasonable toy problem. Thus, is it possible that the linear programming bounds yield something interesting for the coefficient problem on $\{0,1\}^{2^d}$? A direct use does not, simply because we have, in the language of the previous chapter, the following

Lemma 27. *For any n, q and any $c_i \geq 0$ for $1 \leq i \leq n$ with $\sum_i c_i = 1$, $(1, c_1, c_2, \ldots, c_n)$ is a valid quasicode.*

Proof of Lemma 27. Follows immediately from positive definiteness of Krawtchouk polynomials $K_j(0) \geq |K_j(i)|$. □

There are thus numerous possibilities that are perhaps worth exploring for future research. We outline a few here.

1. Is it possible to "shape" the magnitudes of the binary Fourier coefficients arbitrarily in the sense of a multiplicative constant a la Nazarov once we "coalesce" them by Hamming weight? We believe the answer is yes, but do not have a construction at present.

2. It is also possible that the LP bounds are just not enough to give useful information about the coefficient problem. This lack of information from the LP

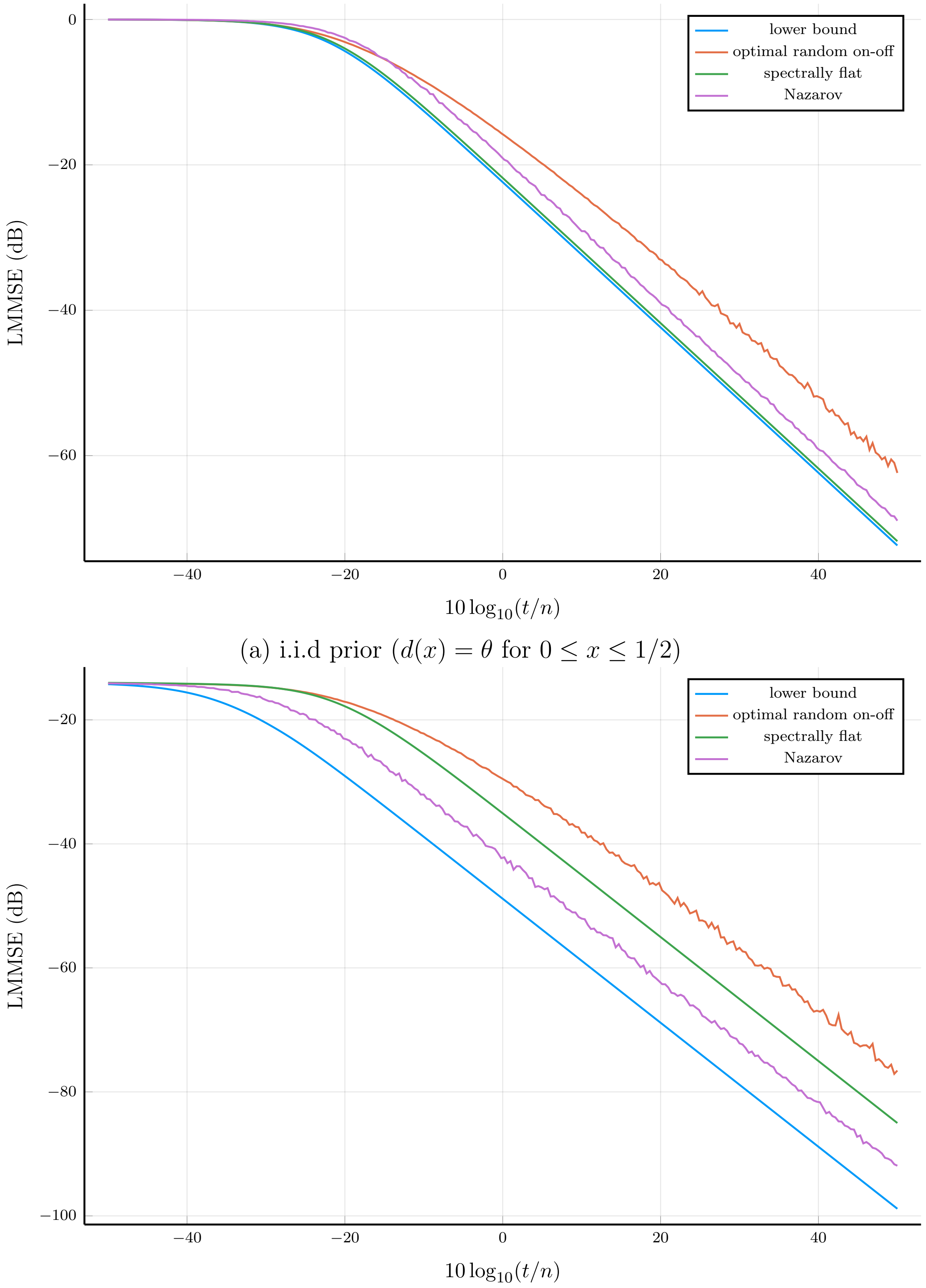
bounds is a mystery for us and we believe this phenomenon worth clarifying further. For example, what happens with higher order SDP bounds and the coefficient problem?

3. Can we use answers to the above to get useful information for other orthonormal bases, such as Fourier on $\mathbb{Z}/2^d\mathbb{Z}$?

Although Prop. 10 shows universal tightness across all priors, even "extreme" ones like bandlimited ones, the constant is worse than that for a spectrally flat construction for i.i.d. scenes. The better performance of spectrally flat constructions over the ones inspired by Nazarov's theorem (in certain regimes) seems to extend to other "natural" priors like the $f^{-\gamma}$ one, as the waterfilling still yields something that is nearly "flat". It might be interesting to quantify and understand the "flatness" of the waterfilling for "natural" priors. Such an analysis is conceptually simple given the contents of the waterfilling Lemma 21 and associated bound Proposition 7.

One issue that we have not addressed here or in [125] is the equal scaling of n at both sensor and scene. One natural way to address this is letting $\mathbf{A}$ be $m \times n$, or alternatively one could study a continuous model. Another issue is obtaining a good understanding of mask/lens combinations. Understanding such combinations will require not only updates to the simple propagation model studied here and in [125], but also a refined understanding of the cost tradeoffs between lenses and apertures.

Stepping back from imaging problems, one may ask the question of where else Nazarov's theorem can be used in applied contexts, something also raised implicitly in [21]. For example, as Nazarov's theorem does not care about orthogonality, but merely a l_2 estimate like Parseval's theorem, one can use it for frames as well as bases, or for anything satisfying a restricted isometry property. Another example is the fact that we merely use the l_∞ case of his theorem which works for all l_p spaces. Furthermore, the astute reader would have noticed that much of the discussion of this chapter relies only on just a few pieces that have a fair bit of slack. We therefore firmly believe that Nazarov's theorem and the (heuristically) efficient greedy algorithm could play interesting roles elsewhere. An entertaining illustration is that of constructing good lattice packings via Bang's lemma [16] as Ball [14] describes. We do not describe Ball's construction here, but simply note that we still agree with Ball's assessment given at the end of his article [14]: "As with other "constructions" of efficient packings, the simplicity here is probably an illusion". Roughly speaking, the construction relies on finding a sign cortège of length at least exponential in n. Without our observation regarding the sufficiency of "local" optimality, this would require at least doubly exponential time in the dimension n. With it, we (heuristically) get rid of one exponentiation to make it singly exponential.



(a) i.i.d prior ($d(x) = \theta$ for $0 \leq x \leq 1/2$)

(b) bandlimited prior ($d(x) = \theta$ for $0 \leq x \leq s - r$, 0 for $x \geq s + r$, and $\theta(s + r - x)/(2r)$ otherwise for $0 \leq x \leq 1/2$)

Figure 3-5: $n = 677, \theta = 1, W = J = 0.001, s = 0.02, r = 0.005$. We use the quartic residue construction for spectrally flat. Jaggedness of the Nazarov plot comes from the fact that in general the spectrum allocation varies with t and we randomly seed the sign cortège.

Chapter 4

New lower bounds for the mean squared error of vector quantizers

> For twelve years I have been studying properties of parallelohedra. I can say it is a thorny field for investigation, and the results which I obtained and set forth in this memoir cost me dear…
> Three-dimensional parallelohedra are now playing an important role in the theory of crystalline bodies, and crystallographers have already paid attention to properties of these strange polyhedra, but till now crystallographers were satisfied with the description of parallelohedra from a purely geometrical point of view. I noticed already long ago that the task of dividing the n-dimensional analytical space into convex congruent polyhedra is closely related to the arithmetic theory of positive quadratic forms.
>
> *Georgy Voronoï*, 1907

4.1 Introduction

Let us shift gears a bit and examine the problem of vector quantization in Euclidean space. Links and interesting connections to material presented in the preceding chapters will become clear as we proceed.

The problem of quantization is of fundamental importance to signal processing, with a long and distinguished history [57]. Our focus in the current chapter is on the mathematical theory. Specifically, we study lower bounds on the mean squared error under the “high-resolution limit”, a study which originated in work on pulse-coded modulation (PCM) [90]. We do note that there is an important complementary perspective offered by rate distortion theory, see e.g. [57] or [92, Ch. 25-27] for more information on this topic and the relation between these two perspectives. The focus on mean squared error is for mathematical simplicity, though even in such a setting fine-grained questions appear difficult. An astute reader will note that a nontrivial amount of the discussion, especially the general setting explored in Theorem 10, is generalizable to distortion measures that go beyond squared error.

It is well known that one can reduce the “high-resolution” problem for general source probability distributions (under weak assumptions [129, Thm. 1]) to that of studying a uniform source over a large region through the use of “companders” in the scalar case [19, 91], and a point density function in the general case [128, 129] [1]. As such, the basic object of study may be defined as follows [30]. For points $p_1, p_2, \ldots, p_M \in [0,1]^d$, define the *normalized second moment (NSM)*, scaled down by a factor of d by

$$G(p_1, \ldots, p_M) \triangleq \frac{1}{d} \frac{\frac{1}{M}\sum_{i=1}^{M} \int_{V(p_i)} |x - p_i|^2 dx}{\left(\frac{1}{M}\sum_{i=1}^{M} |V(p_i)|\right)^{1+\frac{2}{d}}}, \tag{4.1}$$

where $V(p)$ denotes the Voronoï cell associated to p restricted to $[0,1]^d$, and $|\cdot|$ denotes volume. The Voronoï cell $V(p)$ is the set of points closer to p than any other point, that is

$$V(p_i) \triangleq \{x \in \mathbb{R}^d : \forall j \neq i, |x - p_i| < |x - p_j|\}.$$

Note that as defined above, the Voronoï cells do not strictly partition $\mathbb{R}^d$ as the cell boundaries are not included in any of them. Such issues do not concern us at the moment, and play a minimal role in this chapter. One may ignore these issues in our context simply because these boundaries have zero measure, and thus the boundaries

[1] For $d = 2$, this is already implicitly present in [43].

do not affect “bulk” quantities like the NSM. We henceforth reserve the term NSM to refer to the quantity that is not divided by d, so for instance in (4.1), the NSM is $dG(p_1, \ldots, p_M)$. We shall call G itself the *per-dimensional NSM.*

We also find it convenient to talk about the NSM of a body B about a point v, defined by

$$NSM(B, v) \triangleq \frac{\int_B |x-v|^2 dx}{|B|^{1+\frac{2}{d}}}.$$

When the choice of v is clear from context, we may omit it. For example, when we talk of the NSM of a ball, v is implicitly the center of the ball.

Before proceeding further, we comment on the history of Voronoï cells. The original mathematical impetus can arguably be attributed to Gauß, Hermite, and Lagrange, who did a detailed study of quadratic forms in number theory. According to [114, p. 7], Dirichlet delivered an interesting lecture in the physical-mathematical class meeting of the Prussian Military Academy on the 31st of July, 1848 on the reduction of a positive quadratic form with three indeterminate integers. In his successful effort at simplifying the work of Gauß and Seeber on this topic, Dirichlet introduced what we now commonly call a Voronoï cell of a lattice (corresponding to the quadratic form). Hence some authors also call this a “Dirichlet tesselation”. However, Voronoï appears to have been the first to undertake a detailed study of such tessellations in three outstanding papers in Crelle’s journal ([122], [121], [123]). Henceforth we stick to the common practice of talking about Voronoï cells, partitions, decompositions, and tessellations. The middle two are exact synonyms, and the last one, which was the principal object of Voronoï’s study, we reserve for the lattice case only, that is when p_i are elements of a *lattice* $\Lambda \in \mathbb{R}^d$.

Definition 15. *A lattice $\Lambda \in \mathbb{R}^d$ is an additive subgroup of $\mathbb{R}^d$ which is isomorphic to the additive group $\mathbb{Z}^d$, and which spans the real vector space $\mathbb{R}^d$.*

We also note that in applied contexts, the physician and father of modern epidemiology John Snow used a Voronoï partition to illustrate how most people who died in the 1854 Broad Street cholera outbreak lived closer to the infected Broad street pump than to any other pump. For a detailed account of how Snow collected his data and constructed his map, we recommend [68]. Given the simplicity and central importance of this idea, we think it likely that this idea was discussed by many others as well.

Returning to mathematics, we may then define the minimal per-dimensional NSM by

$$G_d \triangleq \lim_{M\to\infty} \inf_{p_i} G(p_1, \ldots, p_M). \tag{4.2}$$

Restricting to the special case where p_i are points of a lattice Λ', (4.1), (4.2) simplify, and one may define a minimal per-dimensional NSM for lattices by

$$G_{\Lambda,d} \triangleq \inf_{\Lambda'} \frac{\int_{V(0)} |x|^2 dx}{d|V(0)|^{1+\frac{2}{d}}}. \tag{4.3}$$

Zador [128, 129] proved

$$G_d \geq \frac{1}{(d+2)\pi} \Gamma\left(\frac{d}{2}+1\right)^{\frac{2}{d}}. \tag{4.4}$$

Zador [128, 129] also obtained an asymptotically matching upper bound

$$G_d \leq \frac{1}{d\pi} \Gamma\left(\frac{d}{2}+1\right)^{\frac{2}{d}} \Gamma\left(1+\frac{2}{d}\right),$$

and thus showed

$$\lim_{d\to\infty} G_d = \frac{1}{2\pi e}.$$

Zador's work [128, 129] contains most of the foundational material on the mathematics of vector quantization, see e.g. the survey of Gray and Neuhoff [57] for details.

Poltyrev [130, Lemma 1] observed that asymptotically good coverings result in asymptotically good quantizers. In particular, one may use the good lattice coverings of Rogers [98] and demonstrate

$$\lim_{d\to\infty} G_{\Lambda,d} = \frac{1}{2\pi e}.$$

4.2 Main results

The main results of this chapter are improvements on (4.4). To describe them, we need some notation. Let

$$V_d \triangleq \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)}$$

be unit ball volume. Then $A_d = dV_d$ is its surface area. Let

$$I_x(a,b) \triangleq \frac{\int_0^x t^{a-1}(1-t)^{b-1}dt}{\int_0^1 t^{a-1}(1-t)^{b-1}dt}$$

be the regularized incomplete beta function. One may compute the solid angle and NSM of a circular solid cone of diagonal 1, height h about its vertex and obtain

$$\begin{aligned}\Theta(d,h) &= \frac{1}{2}A_d I_{1-h^2}\left(\frac{d-1}{2},\frac{1}{2}\right),\\ \xi(d,h) &= \frac{d^{1+\frac{2}{d}}}{d+2}\frac{h^2+(1-h^2)\frac{d-1}{d+1}}{(hV_{d-1}(1-h^2)^{\frac{d-1}{2}})^{\frac{2}{d}}}.\end{aligned}$$

We then have the following conjecture:

Conjecture 2. *Let $d \geq 2$. Let $F_d \triangleq 2(2^d-1)$, and let h_d be the root of $\Theta(d,h) - \frac{A_d}{F_d} = 0$. Then,*

$$G_{\Lambda,d} \geq \frac{\xi(d,h_d)}{dF_d^{\frac{2}{d}}}. \tag{4.5}$$

Taking

$$F_{d,k} \triangleq 2(2^d-1)+(k-1)2^d \tag{4.6}$$

in the above bound as opposed to F_d, one obtains a lower bound on G_d restricted to quantizers formed by k translates of a lattice.

We believe that we have firm evidence in favor of this Conjecture 2. The missing ingredients can be captured by certain technical inequalities that may be readily verified on a computer (such as (4.19)), but that we are unfortunately unable to rigorously prove.

One of the chief aims of this chapter is to enable the reader to understand where this conjecture comes from, and why we view it as extremely plausible. We also believe that Conjecture 2 should be easier to prove than Conway and Sloane's conjectured bound [31], though we note that Conway and Sloane's conjectured bound is sharper and applies to all quantizers, not just lattice quantizers.

We are sympathetic to the reader who is dissatisfied with the state of affairs regarding the missing technicalities required to prove Conjecture 2. Such readers may find solace in the following rigorous improved lower bound valid for all quantizers, and not just lattice ones:

Theorem 10. *Define $\nu(d,r)$ for dimension $d \geq 1$, $0 \leq r \leq 1$ to be the NSM of a truncated unit ball centered at the origin and computed about the origin. Here the ball is truncated by intersecting it with a hyperplane at distance r from the origin, such as $x_1 \leq r$. At the limit $r = 0$, we have a hemisphere, and at $r = 1$, we have the original unit ball. Define*

$$\kappa(d,r) \triangleq \min_{0 \leq x \leq r} \nu(d,r).$$

Let $c_d = \kappa(d,1)$ be the NSM of the unit ball, namely

$$c_d \triangleq \frac{d}{(d+2)\pi}\Gamma\left(\frac{d}{2}+1\right)^{\frac{2}{d}}.$$

Let $\gamma_d \triangleq 1 - \frac{3}{2}\left(\frac{2}{3}\right)^d$, and let k_p be the base of the exponent of the asymptotic sphere packing bound of Kabatyanskii and Levenshtein [69], given by

$$k_p = 2^{-0.599\ldots} \approx 0.6602.$$

Let $d \geq 3$ be sufficiently large. Let $f(z)$ be an n-level quantizer on $[0,1]^d$. Then,

$$\int_{[0,1]^d} |z - f(z)|^2 dz \geq \max\left(n^{-\frac{2}{d}} c_d \frac{1}{2}\left(\left(\frac{2}{3}\right)^{1+\frac{2}{d}} + \left(\frac{4}{3}\right)^{1+\frac{2}{d}}\right),\right.$$
$$\left. n^{-\frac{2}{d}}\left(\left(\gamma_d - \frac{1}{2}\right)\kappa\left(d, \frac{3}{2}k_p\right)^{-\frac{d}{2}} + \left(\frac{3}{2} - \gamma_d\right) c_d^{-\frac{d}{2}}\right)^{-\frac{2}{d}} + o(n^{-\frac{2}{d}})\right),$$

as $n \to \infty$.

To the best of our knowledge, Theorem 10 represents the first rigorous improvement over the bound of Zador (4.4).

4.3 Proofs

At a high level, our strategy for lattice quantizers may be described as follows. The original "sphere bound" of Zador [128], [129] comes from the fact that each Voronoï cell's second moment, normalized by its volume, can't be lower than that of a ball. This trivially provable fact is very nice as it immediately yields Zador's asymptotically tight bound upon carrying out the computation. Our approach here is heavily inspired by the work of Tóth/Newman [116], [88], who prove a sharp bound for $d = 2$ by utilizing an upper bound on the number of edges of the Voronoï polygons. This upper bound on the number of edges, together with some calculus, convexity, and Hölder's inequality allows one to prove that the hexagonal lattice quantizer is optimal for $d = 2$. What we do is simply work with upper bounds on facet counts in higher dimensions and appropriately generalize the machinery. There is one serious drawback of this approach: there is a finite upper bound on facet counts only for the lattice case for $d \geq 3$, see e.g. work of Dolbilin and Tanemura [39, §5] for a construction for $d = 3$ of arbitrarily large facet counts for a Voronoï cell in a non-lattice quantizer. Our methods therefore bifurcate into separate lattice and general quantizer cases.

Although we find the above approach to lattice quantizers more interesting as compared to our methods for the general case, we shall first study the general case and prove Theorem 10. Our justification for this ordering is that it illustrates some of the basic convexity and Hölder's inequality machinery that plays a key role in the lattice case as well. In fact, even before studying the general case, we give a brief rederivation of Zador's lower bound (4.4) using convexity. We have not found this approach in the literature.

4.3.1 Rederivation of Zador's lower bound

Let us for simplicity look at lattice quantizers Λ with $|\Lambda| = 1$ in $\mathbb{R}^d$. By $|\Lambda| = 1$ we mean that the volume of a fundamental cell of the lattice is 1. We have the basic

$$\forall x, \forall \beta > 0, \quad \min_{v \in \Lambda} |x-v|^2 \geq -\frac{1}{\beta} \log \left(\sum_{v \in \Lambda} e^{-\beta |x-v|^2} \right) \tag{4.7}$$

Let the quantizer value (second moment) be denoted $NSM(\Lambda)$; we have ensured normalization by the assumption that $|\Lambda| = 1$. We have by the concavity of log and (4.7)

$$\begin{aligned} NSM(\Lambda) &= \int_{x \in V(0)} \min_{v \in \Lambda} |x-v|^2 dx \\ &\geq -\frac{1}{\beta} \int_{x \in V(0)} \log \left(\sum_{v \in \Lambda} e^{-\beta |x-v|^2} \right) dx \\ &\geq -\frac{1}{\beta} \log \left(\int_{x \in V(0)} \sum_{v \in \Lambda} e^{-\beta |x-v|^2} dx \right) \\ &= d \left(\frac{1}{2\pi} \frac{\log \left(\frac{\beta}{\pi} \right)}{\frac{\beta}{\pi}} \right). \end{aligned}$$

Optimizing over β, one picks $\beta = \pi e$. This choice of β yields $NSM(\Lambda) \geq \frac{d}{2\pi e}$, which is in fact the asymptotic value ($d \to \infty$) of the minimum NSM, as proved by Zador [129]. Can we do better and actually recover the non-asymptotic lower bound of Zador (4.4)? It turns out that we can!

More abstractly, let us consider what we need for the above argument. Let us consider a radial function $f(r^2)$ satisfying:

$$f \geq 0, f' \leq 0, f'' \geq 0.$$

We also certainly want $f \in l_1(\mathbb{R}^d)$, and the inverse to make sense on the infinite sum $\sum_{v \in \Lambda} f(|x-v|^2)$. With these conditions met, we may replace $e^{-\beta r^2}$ by f.

It may be checked that the exact optimal f meeting these conditions is a "tent" function of the appropriate scale, where a "tent" function is of the form $f(r) =$

$(a - br)^+$ with $x^+ = \max(x, 0)$ and $a, b > 0$. This choice of f may be justified as follows. The tent functions are the extremal rays of the cone given above, and it is easily checked that using a convex combination of functions for the above argument does no better than the best extremal endpoint. All that remains is to pick the right scale for the tent. Upon optimizing the scale of the tent and performing the relevant computation, one recovers Zador's non-asymptotic lower bound (4.4).

The above argument generalizes to quantizers formed by k translates of a lattice, where k is a finite number. Such quantizers can come arbitrarily close in performance to optimal quantizers by standard limiting arguments. For example, just consider quantizing the unit cube at arbitrarily fine resolution, and then replicating the quantization points contained in the unit cube across space by translating them by $\mathbb{Z}^d$. Thus in fact the argument of this section is a genuine rederivation of Zador's lower bound (4.4).

4.3.2 The general case

Hopefully the reader is now convinced of the value of convexity considerations in the study of the minimal NSM of quantizers. We now perform a more detailed study and prove Theorem 10.

We start off with a basic inequality.

Lemma 28. *Let $a_i, b_i \geq 0$, $\sum_i a_i = 1$, $p > 1$, and let $b_i \geq c > 0$ for all i. Then, we have*

$$\sum_i a_i^p b_i \geq \max\left(c \sum_i a_i^p, \left(\sum_i b_i^{\frac{1}{1-p}}\right)^{1-p}\right).$$

Proof of Lemma 28. The first term in the maximum is trivial, and the second term is immediate from Hölder's inequality. □

In our application of Lemma 28, c will be the NSM of a ball, b_i will be the NSM of Voronoï cells, a_i will be the volumes of the Voronoï cells in a decomposition of the unit square, and finally $p = 1 + 2/d$ where d is the dimension of the vector quantizer.

Plugging in $b_i \geq c$ into Lemma 28, and dropping the first term out of the maximum immediately yields Zador's "sphere bound" (4.4) once the number of points $n \to \infty$ as we defined earlier. Plugging in $a_i = 1/n$ and dropping the second term also yields Zador's sphere bound. Thus our goal is to somehow get a nontrivial trade-off between the two terms. Quantifying this trade-off amounts to the task of showing that at least one of the following phenomena must take place for any vector quantizer:

1. A nontrivial amount of "dispersion" in the volumes of the Voronoï cells.

2. A nontrivial fraction of the b_i are bounded away from the ball's NSM.

Let us first understand why one would might expect this. Consider the "extreme" case for the first item, which happens with lattice quantizers where all Voronoï cells are identical. Each point of the lattice has a "close" nearest neighbor since lattices can't have packing density exceeding standard *sphere packing density* bounds. Heuristically, a packing of spheres is just a non-overlapping collection of equally sized balls in Euclidean space. Its density is given by the limiting ratio of the volume occupied by the balls to the volume of a large bounded region, with the limit taken as the region grows to infinity. As we are not proving statements about sphere packing here, we simply refer the reader to e.g. [30, Ch. 1] for a rigorous definition of sphere packing density. The current best known (asymptotic) upper bound is the classical $2^{(-0.599\cdots+o(1))d}$ due to Kabatyanskii and Levenshtein [69], who used Delsarte's linear programming method [35] together with a suitably modified version (for the Euclidean sphere) of the construction of McEliece, Rodemich, Rumsey, and Welch (MRRW) [82] done for Hamming space. The sphere packing bound implies that the b_i must be bounded away from c: at best their NSM matches that of a ball cut off by a hyperplane governed by the packing radius. The sphere packing bounds apply not just to lattices, but also arbitrary point configurations! We may therefore proceed further.

To mathematically quantify this phenomenon, the following is useful.

Lemma 29. *Let $x, y \geq 0$, and $p > 1$, and let $w, a, b \in (0, 1)$. Consider the optimiza-*

tion problem

$$\begin{aligned} \min F &\triangleq wx^p + (1-w)y^p, \\ \textit{subject to } wx + (1-w)y &= 1, \\ x &\leq a, \\ w &\geq b. \end{aligned}$$

Then the minimum is attained at $x = a, w = b$.

Proof of Lemma 29. First, let w be fixed, and consider optimizing over x. Setting $y = \frac{1-wx}{1-w}$, we get

$$\frac{\partial F}{\partial x} = -pw\left(\left(\frac{1-wx}{1-w}\right)^{p-1} - x^{p-1}\right).$$

As $x < 1, p > 1$, we see that F is decreasing in x. Thus for a fixed w, we must set $x = a$ to minimize F.

Now, setting $x = a$, we claim that F is increasing in w. Showing this would complete the proof. Once again, studying $\frac{\partial F}{\partial w}$, we reduce our task to showing:

$$(1-aw)^{p-1}[1 - aw - (1-a)p] \leq (a - aw)^p. \tag{4.8}$$

We have equality at $a = 1$, suggesting the following proof of (4.8). We may rewrite (4.8) as

$$(1-aw)^p - (a-aw)^p \leq (1-a)p(1-aw)^{p-1}.$$

But observe that x^p is convex since $p > 1$, and so we have:

$$(1-aw)^p - (a-aw)^p \leq [(1-aw) - (a-aw)]p(1-aw)^{p-1},$$

as desired. □

We may now proceed with the Proof of Theorem 10 along the sketched direction above.

Proof of Theorem 10. Let the points of the quantizer be $\lambda_1, \lambda_2, \dots, \lambda_n$. Let the minimum distance for each of these points to their nearest neighbors be denoted by r_i. Define $r_d(v)$ to be the radius of the ball in d dimensions of volume v. By the sphere packing consideration, we see that for any $0 < \gamma < 1$, and sufficiently large d, $\gamma n + o(n)$ of the λ_i must have

$$r_i \leq \left(\frac{1}{1-\gamma}\right)^{\frac{1}{d}} k_p r_d \left(\frac{1}{n}\right). \tag{4.9}$$

For if this statement was not true, we could restrict our attention to the $(1-\gamma)$ fraction of points with largest minimum distance and violate the sphere packing density upper bound.

Here k_p is the base of the exponent in an asymptotic upper bound for sphere packing density of the form $\Delta_d \leq k_p^{d+o(d)}$, and we may take by [69]

$$k_p = 2^{-0.599\dots} < \frac{2}{3} < 1.$$

We emphasize that the argument does not rely on the precise numerics of k_p, though our choice of certain parameters here does. For example, at a certain step we chose $2/3$ simply because $(3/2)k_p < 1$ and it is a simple fraction. All one really needs is $k_p < 1$. For a reader interested in elementary arguments that still yield a nontrivial k_p, we recommend the work of Blichfeldt [20] who obtained

$$\Delta_d \leq \frac{d+2}{2}\left(\sqrt{\frac{1}{2}}\right)^d,$$

valid for all $d \geq 1$.

Now let f denote the fraction of the number of points λ_i with associated Voronoï cell volumes $|V(\lambda_i)| \geq 2/(3n)$. Now we divide into two cases depending on the value of f.

1. Suppose $f \leq 1/2$. Here we have sufficient “dispersion” in the volumes of the

Voronoï cells. We may apply Lemma 29 to immediately get a lower bound

$$\int_{[0,1]^d} |z - f(z)|^2 \geq n^{\frac{-2}{d}} c_d \frac{1}{2} \left(\left(\frac{2}{3}\right)^{1+\frac{2}{d}} + \left(\frac{4}{3}\right)^{1+\frac{2}{d}} \right), \tag{4.10}$$

where c_d is the NSM of the ball. This covers the first term of our basic approach outlined in Lemma 28.

2. Now suppose $f > 1/2$. Here we no longer rely on "dispersion" of the volumes, but instead use the fact that we have a nontrivial fraction of sufficiently "large" Voronoï cells. We also know by (4.9) that a large fraction of points have short minimum distance. These two sets must have a nontrivial intersection.

   Mathematically, we need to choose γ appropriately. What we have is:

   $$r_i \leq r_d \left(\frac{1}{1-\gamma} k_p^d \frac{1}{n} \right),$$

   and we will make the argument of r_d match $[2/(3n)] \left((3/2)k_p\right)^d$ for example. We chose $3/2$ simply because $(3/2)k_p < 1$. We therefore take

   $$\gamma_d \triangleq 1 - \frac{3}{2} \left(\frac{2}{3}\right)^d \geq \frac{5}{9},$$

   since we may take $d \geq 3$. For sufficiently large d, we thus get

   $$\int_{[0,1]^d} |z-f(z)|^2 \geq n^{-\frac{2}{d}} \left((\gamma_d + \frac{1}{2} - 1)\kappa \left(d, \frac{3}{2}k_p\right)^{-\frac{d}{2}} + \left(\frac{3}{2} - \gamma_d\right) c_d^{-\frac{d}{2}} \right)^{-\frac{2}{d}} + o(n^{-\frac{2}{d}}), \tag{4.11}$$

   where the little o is with respect to n and d is fixed.

Combining (4.10) and (4.11) gives us Theorem 10. □

Remark 9. *No attempt has been made to optimize the parameters of Theorem 10. Part of the difficulty lies in understanding quantities like $\kappa(d, r)$. The astute reader may, for example, conjecture that $\kappa(d, r) = \nu(d, r)$, which would follow from the natural conjecture that $\nu(d, r)$ is decreasing in r for $r \in [0, 1]$. We do not have a proof* *of this statement. Indeed, more generally, we do not have a good grasp on the NSM of objects derived in natural fashions from balls, such as the $\xi(d,h)$ of Conjecture 2. This difficulty extends to our subsequent discussion in subsection 4.3.4 on lattice quantizers as well. It is also clear that d being sufficiently large can be replaced by $d \geq 2$ as long as one uses a non-asymptotic sphere packing bound.*

Let us now turn to the study of lattice quantizers, where we generalize the work of Tóth/Newman on the optimality of the hexagonal lattice quantizer for $d = 2$ among all quantizers, and not just lattice ones!

We believe that the best way to understand our generalization of is by first examining the original work of Tóth/Newman [116], [88]. More precisely, we follow Newman's presentation essentially verbatim as it is both mathematically simpler and also available readily in English. We do not know of any essential simplification to his argument, presented in the following subsection 4.3.3.

4.3.3 Tóth/Newman's hexagon theorem

Theorem 11. *Let S be the unit square $[0,1]^2$, and $f(Z)$ any function on S taking at most n distinct values, i.e. an n-level quantizer. Then,*

$$\int_S |z - f(z)|^2 dz > \frac{\sigma}{n}, \text{ where } \sigma = \frac{5}{18\sqrt{3}}.$$

Recall that Voronoï cells form a partition of S, bounded by straight line segments. We prove a Lemma regarding the number of edges formed by any partition of S into convex polygons (not necessarily Voronoï!) as a direct consequence of Euler's theorem on planar graphs:

Lemma 30. *Let S be divided into k convex polygons in any manner, and let E denote the total number of edges. Then $E \leq 3k + 1$.*

Proof of Lemma 30. At each vertex v we let m_v be the number of edges meeting at it. By double counting, $2E = \sum_v m_v$. Furthermore, $m_v \geq 3$ at all v except possibly at the corners of the square S due to convexity of the polygons. Thus,

$$2E = \sum_v m_v \leq 2 \times 4 + 3 \left[\sum_v (m_v - 2) \right] = 6E - 6V + 8,$$

where V is the total number of vertices. By Euler's theorem, $E - V = k - 1$, so the proof is complete. □

We now apply Lemma 30 to a slightly modified Voronoï diagram, where we partition the Voronoï cells further into right and obtuse angled triangles according to a certain procedure described below and prove the following Lemma.

Lemma 31. *Let λ_i, $1 \leq i \leq n$ denote the quantizer outputs. Let $V(\lambda_i)$, $1 \leq i \leq n$ denote their respective Voronoï cells. For each $V(\lambda_i)$, draw in it all the line segments from λ_i to the vertices of $V(\lambda_i)$. Also draw in all the perpendiculars from λ_i to the edges of $V(\lambda_i)$, possibly extended, and terminate these at the boundary of $V(\lambda_i)$. This construction procedure subdivides the Voronoï cells into triangles, each of which has a right or obtuse angle.*

On performing this procedure over all the $V(\lambda_i)$, we end up with triangles $T_1, T_2, \ldots, T_N$, and vertex angles formed at the corresponding λ_i that we call $\theta_1, \theta_2, \ldots, \theta_N$. Then,

$$N \leq 12n - 4.$$

Proof of Lemma 31. Let N_i be the number of triangles T_k formed through the subdivision process given above on $V(\lambda_i)$. Let E_i be the number of edges of $V(\lambda_i)$. Once again, by a double counting argument (taking care of the boundary of S), we have

$$\sum_i E_i = 2E - (\text{number of edges on boundary of } S) \leq 2E - 4.$$

But then

$$N = \sum_i N_i = 2 \sum_i E_i \leq 4E - 8 \leq 12n - 4,$$

where in the last inequality we used Lemma 30. □

We now show that the NSM of a triangle T_k about its vertex is lower bounded by that of a right angled triangle with the same vertex angle θ_k.

Lemma 32. *Define*

$$\phi(\theta) \triangleq \frac{3\tan(\theta)}{3+\tan^2(\theta)}. \tag{4.12}$$

If T is any right or obtuse triangle and λ is a vertex of T with acute vertex angle θ, then

$$\int_T |z-\lambda|^2 dz \geq \frac{|T|^2}{\phi(\theta)}.$$

Proof of Lemma 32. The proof is essentially by a geometric comparison with a suitable right angled triangle with the same vertex, same vertex angle, and same area. The details are given below.

Let e be the edge of T whose endpoints are λ and the right or obtuse vertex of T, let f be the other edge of T containing λ, and finally let g be the third edge. From some point x on f, we drop a perpendicular to e (possibly extended) and call the foot of the perpendicular y. We choose x such that the right angled triangle formed by λ, x, y has the same area as T; this is what we mean by the "geometric comparison" alluded to above. Let us call this right angled triangle U. Let the intersection of xy with g be z_0.

It is clear that

$$\begin{aligned} \forall z \in T-U, &\quad |z-\lambda| \geq |z_0-\lambda|, \\ \forall z \in U-T, &\quad |z-\lambda| \leq |z_0-\lambda|. \end{aligned}$$

Thus, with the integrand $|z-\lambda|^2$ and the implicit Lebesgue measure, we have

$$\int_T = \int_{T\cap U} + \int_{T-U} \geq \int_{T\cap U} + \int_{U-T} = \int_U = \frac{|U|^2}{\phi(\theta)},$$

where for the last equality one performs a direct calculation and is indeed how one comes across ϕ in the first place. The proof is complete since by construction $|U| = |T|$. □

We now explicitly note that ϕ is concave:

Lemma 33. *ϕ governed by (4.12) is concave on $\left(0, \frac{\pi}{2}\right)$.*

Proof of Lemma 33. One may readily calculate

$$\phi''(\theta) = \frac{-48\sin^3(\theta)\cos(\theta)}{(\cos(2\theta)+2)^3},$$

and the nonpositivity on $(0, \pi/2)$ is obvious. $\square$

Remark 10. *We are not satisfied with the proof of Lemma 33 above as it is not conceptual enough and relies on a direct calculation. The reader may not view this as a defect right now since the differentiation and associated concavity check is easily observed above. However, in our efforts to generalize this method to $d \geq 3$, we run into a nontrivial assertion of concavity that we are unable to prove (Conjecture 4), but can simulate to whatever precision we want on a computer. The nontriviality of Conjecture 4 is ultimately why we must still leave Conjecture 2 as a Conjecture, and not a Theorem. We note that neither Newman [88] (who asserted the concavity of ϕ as defined above) nor Tóth [116] (who had a slightly different ϕ but relied on concavity as well) have a more elegant proof. We therefore consider it worthwhile to find an alternative proof.*

All the pieces are now in play to prove the "hexagon" Theorem 11:

Proof of Theorem 11. Consider the equation $\phi(\theta) = \frac{\theta}{2\pi\sigma}$. Equality holds at $\theta = 0$ and $\theta = \frac{\pi}{6}$, so by concavity, we immediately get

$$\phi(\theta) < \frac{\theta}{2\pi\sigma} \tag{4.13}$$

on $\left(\frac{\pi}{6}, \frac{\pi}{2}\right)$. We may thus write, by Jensen's inequality,

$$\sum_{k=1}^{N} \phi(\theta_k) \leq N\phi\left(\frac{\sum \theta_k}{N}\right) = N\phi\left(\frac{2\pi n}{N}\right) < N\frac{n}{\sigma N} = \frac{n}{\sigma}. \tag{4.14}$$

In the second inequality we used Lemma 31 to write $N < 12n$, and we combined this with the above (4.13).

We complete the proof by an application of Cauchy-Schwarz. Using the bound (4.14), we have

$$\int_S |z - f(z)|^2 \geq \sum_{k=1}^{N} \frac{|T_k|^2}{\phi(\theta_k)} \geq \frac{(\sum |T_k|)^2}{\sum \phi(\theta_k)} \geq \frac{\sigma}{n}.$$

□

4.3.4 Conjectured bound for lattices

As remarked upon earlier, our approach to improved lower bounds for lattice quantizers is best described as a generalization of the preceding subsection 4.3.3. Our thinking may be described as follows:

1. We can not rely on Euler's theorem to get upper bounds on facet counts for $d \geq 3$. Our solution is to focus in on lattice quantizers only, and rely on the original work of Minkowski [84, §6] and independently Voronoï [121, §48] for an upper bound on the facet counts in this case. We note that one can generalize this to a union of k translates of a lattice quantizer by the work of Delone and Sandakova [34], indeed this is where the $F_{d,k}$ of (4.6) comes from.

2. We need to think of a relevant decomposition of Voronoï polyhedra to which we can apply the concavity considerations described in the proof of Theorem 11. There are two natural candidates.

   One is closer in spirit to the work of Newman [88], and decomposes the Voronoï polyhedra into orthosimplices. Here, one needs upper bounds not just on facet counts, but also lower dimensional faces of the polytope. These are also due to Voronoï [121, §65], and we give these bounds in a remark later 11. The decomposition into orthosimplices was our first attempt. However, the technical difficulties of coming up with a good lower bound on the second moment with respect to the vertex of the orthosimplex, subject to fixed solid angle (as opposed to regular angle) and analogous to Lemma 32, defeated us. More importantly, even if we could resolve these difficulties, we believe that the conjectured bound following such an approach is weaker than that of the second candidate described below.

Our second candidate is closer in spirit to the earlier work of Tóth [116]. Here, we do not construct perpendiculars, but simply use the edges connecting the quantizer points to the vertices of the Voronoï cells. We do not have orthosimplices, but rather pyramids on top of each of the facets of the Voronoï cells for $d \geq 3$. In Tóth's case for $d = 2$, these are triangles. Tóth showed by a geometric comparison inequality similar in spirit to Newman's above in the proof of Lemma 32 that one could lower bound the NSM about the vertex by that of an isosceles triangle with the same vertex angle. Our candidate for the appropriate generalization of the isosceles triangle to $d \geq 3$ is a right circular cone of the appropriate solid angle (Conjecture 3). Here, we are able to obtain much more significant progress, such as a reduction of this Conjecture 3 to a technical inequality given in Conjecture 5. However, even if we could prove Conjecture 5, we are still currently defeated by the analog of the concavity of ϕ (Lemma 33) for $d \geq 3$, expressed as Conjecture 4. The rest of the machinery carries through naturally, with Cauchy-Schwarz replaced by Hölder's inequality. The end result of this reasoning is our conjectured lattice NSM lower bound given by Conjecture 2.

Let us now spell out some of the details. We emphasize the conceptual aspects and leave the missing technicalities as conjectures (specifically Conjectures 3 and 4). In the subsequent subsection 4.3.5, we detail some nontrivial progress towards these conjectures and also justify why they may be verified readily on a computer even if we do not have a proof yet.

For lattices, one may examine a single Voronoï cell. By examining $\mathbb{Z}^d \mod 2$ component-wise, Minkowski [84, §6] and Voronoï [121, §48] independently proved

Lemma 34 (Minkowski/Voronoï upper bound on facet count). *For a lattice Λ, the* *number of facets of a Voronoï cell $F(\Lambda)$ satisfies*

$$F(\Lambda) \leq F_d \triangleq 2(2^d - 1).$$

In order to prove Lemma 34, we have

Definition 16. *A* relevant vector *v for a lattice Λ is a vector in Λ such that v is a normal for one of the faces of the Voronoï cell $V(0)$.*

We now prove a Lemma that characterizes relevant vectors. One can consult the originals [84], [121] for a proof. Instead we follow the exposition of the far more readily available book by Conway and Sloane [30, Thm. 10].

Lemma 35 (Minkowski/Voronoï characterization of relevant vectors)**.** *A nonzero vector $v \in \Lambda$ is relevant iff $\pm v$ are the only shortest vectors in the coset $v + 2\Lambda$. Here, 2Λ is the dilation of Λ by the factor 2.*

Proof of Lemma 35. Note that every nonzero $v \in \Lambda$ determines a halfspace

$$H_v \triangleq \{x \in \mathbb{R}^d : \langle x, v\rangle \leq \frac{1}{2}\langle v, v\rangle\},$$

and the Voronoï cell $V(0)$ is the intersection of all H_v. In fact it is the intersection of just the H_v where v is relevant by the definition of relevance.

Let us prove the "only if" direction first. Suppose $v \equiv w \pmod{2\Lambda}$, $v \neq \pm w$, and suppose without loss of generality that $|w| \leq |v|$. Then $t = (v+w)/2, u = (v-w)/2$ are also nonzero vectors in Λ. Now, if $x \in H_t \cap H_u$, then we have $\langle x, t\rangle \leq (1/2)\langle t, t\rangle, \langle x, u\rangle \leq (1/2)\langle u, u\rangle$. Adding, expanding, and using $\langle w, w\rangle \leq \langle v, v\rangle$, we get $\langle x, v\rangle \leq (1/2)\langle v, v\rangle$. Thus H_v is not needed to define the cell $V(0)$, and so v is not relevant.

Now let us prove the "if" part. Suppose v is not relevant. Then, $v/2$ must lie on or outside some H_w for a nonzero $w \neq v \in \Lambda$. In other words, $\langle v, w\rangle \geq \langle w, w\rangle$. This expression can be rewritten as $|v - 2w|^2 \leq |v|^2$. But $v - 2w \neq \pm v$, and it is also in the coset $v + 2\Lambda$. □

Using Lemma 35, we may easily complete the proof of Lemma 34:

Proof of Lemma 34. Let $v_1, v_2, \ldots, v_d$ be a basis for Λ. Then, every $v \in \Lambda$ can be associated with a vector $(a_1, a_2, \ldots, a_d) \in \mathbb{Z}^d$ by its basis expansion $v = \sum_{i=1}^{d} a_i v_i$. By the above Lemma 35, we see that we can have at most two relevant vectors (corresponding to the sign choice in the ±) to each nonzero $\pmod 2$ residue class. There are $2^d - 1$ nonzero residue classes in $\mathbb{Z}^d$, so we get the desired F_d upper bound on the face counts. □

Remark 11. *For a proof and generalization of the facet count upper bound to a union of k translates of Λ in English, see e.g., [39, p.181-183]. Alternatively, one can study the original work of Delone and Sandakova [34]. Another interesting generalization may be found in Voronoï's original [121, §65], where he proves the upper bound*

$$K_{v,d} \leq (d+1-v)\Delta^{d-v}(m^d)|_{m=1},$$

where $K_{v,d}$ is the number of faces of dimension v for a Voronoï cell of a lattice $\Lambda \subset \mathbb{R}^d$ and $\Delta(f(x)) = f(x+1) - f(x)$ is the finite difference operator. Taking $v = d-1$ we recover the bound on the number of facets, which are nothing but $d-1$-dimensional faces.

As an illustration of Lemma 34, consider $d = 2$ where we get an upper bound of 6 on the number of sides of Voronoï polygons of a lattice. A very interesting aspect that we saw earlier in Lemma 31 is that this upper bound of 6 carries over in the "averaged" sense to non-lattice 2-dimensional vector quantizers, via an application of Euler's theorem for planar graphs. This aspect of Lemma 31 is the basic reason why Tóth/Newman [116], [88] works as a bound for all quantizers for $d = 2$, not just lattice ones. We lose this generality when we go to $d \geq 3$.

The next ingredient we need is our hypothesis that right circular cones have minimal NSM about a vertex subject to a solid angle constraint.

Conjecture 3. *Let H be a fixed hyperplane not passing through 0 and $B \subset H$ a compact, convex set. Let the solid angle of the pyramid formed by B and the vertex*

0 *about* 0 *be* Θ. *Then, the NSM of this pyramid is lower bounded by that of a right circular cone with vertex* 0 *and the same solid angle* Θ.

Let us denote the NSM of a right circular cone with solid angle Θ (in d dimensions) $\phi_d(\Theta)^{-1}$, in analogy with Newman's argument. As in Newman's argument, we then have a concavity hypothesis.

Conjecture 4. $\phi_d(\Theta)^{\frac{d}{2}}$ *is concave on* $(0, A_d/2)$.

Assuming the veracity of Conjecture 3, we may proceed further with Newman's argument by replacing Cauchy-Schwarz with Hölder's inequality and obtain

$$\left(\sum_{k=1}^{N} \frac{|B_k|^{1+\frac{2}{d}}}{\phi_d(\Theta_k)}\right)^{\frac{d}{d+2}} \left(\sum_{k=1}^{N} \phi_d(\Theta_k)^{\frac{d}{2}}\right)^{\frac{2}{d+2}} \geq 1.$$

Here, we are partitioning the Voronoï cells $V(\lambda_1), V(\lambda_2), \ldots, V(\lambda_n)$ for the n point quantizer into pyramids formed by the λ_i and the $d-1$–dimensional facets of $V(\lambda_i)$, and call these $B_k, 1 \leq k \leq N$. Θ_k are the respective solid angles of B_k with their associated λ_j. The analog of $N \leq 12n - 4$ for Newman's argument (Lemma 31) is Lemma 34 which shows $N \leq 2(2^d - 1)n + o(n)$. At $d = 2$, we now have a coefficient of 6 in front of n as opposed to 12, this is precisely because our partitioning is different from that of Newman and is in fact more closely aligned with that of Tóth. The $o(n)$ comes simply from boundary effects at the edge of the unit cube $[0,1]^d$.

Finally, assuming the veracity of the concavity hypothesis (Conjecture 4), we may then complete the argument for Conjecture 2.

4.3.5 Towards a proof for lattices

At this stage, we hope that the reader is convinced that the beef of our approach towards lattices lies in establishing Conjectures 3 and 4. We now perform certain reductions that may assist in proving these conjectures and may clarify certain conceptual aspects. At the very least, they provide a method for computer verification.

Let us draw the normal n to the hyperplane H passing through the vertex 0, and parametrize the convex body B by a function $f(\theta)$, $0 \leq \theta < \frac{\pi}{2}$, $0 \leq f \leq 1$, and f

nonincreasing with θ. Here, θ describes the circular locus of points H_θ lying on H forming angle θ with the normal n. $f(\theta)$ denotes the relative measure of the body at that angle, given by

$$f(\theta) = \frac{|H_\theta \cap B|}{|H_\theta|}.$$

The above definition makes it clear why $0 \leq f \leq 1$. The fact that f is nonincreasing follows at once from the convexity of B and the that we may assume without loss of generality that the foot of n on H lies inside B. For if the foot was outside B, we may translate B so that its center of mass is the foot of the perpendicular and thereby not increase the NSM.

Let us now write down various quantities of interest in terms of f.

First, the solid angle Θ formed at the vertex is

$$\Theta = \int_0^{\frac{\pi}{2}} f(\theta) A_{d-1} \sin^{d-2}(\theta) d\theta. \tag{4.15}$$

Next, the volume of the pyramid V is

$$V = \frac{d-1}{d} \int_0^{\frac{\pi}{2}} f(\theta) \tan^{d-2}(\theta) \sec^2(\theta) d\theta. \tag{4.16}$$

And finally, the second moment M is

$$M = \frac{d-1}{d+2} \int_0^{\frac{\pi}{2}} f(\theta) \tan^{d-2}(\theta) \sec^4(\theta) d\theta. \tag{4.17}$$

Observe that M, V are linear functionals in f. Let us examine

$$NSM^{-\frac{d}{2}} = \frac{V^{\frac{d+2}{2}}}{M^{\frac{d}{2}}}.$$

By Hölder's inequality, it is immediately clear that

$$\frac{\left(tV_1 + \bar{t}V_2\right)^{\frac{d+2}{2}}}{\left(tM_1 + \bar{t}M_2\right)^{\frac{d}{2}}} \leq t\frac{V_1^{\frac{d+2}{2}}}{M_1^{\frac{d}{2}}} + \bar{t}\frac{V_2^{\frac{d+2}{2}}}{M_2^{\frac{d}{2}}},$$

where $t \in [0,1], \bar{t} = 1 - t$. Thus, $NSM^{-\frac{d}{2}}$ is a convex function in f, and we are interested in maximizing it subject to fixed solid angle. Now note that the constraints on f (namely $0 \leq f \leq 1$, $\int f(\theta) d\mu(\theta) = c$ for μ given by (4.15), and f nonincreasing) are all convex constraints, and thus we are maximizing a convex function over a compact, convex, set. We may apply the classical Krein-Milman Theorem [73] to therefore reduce the study to the extremal f. We may focus on step functions without loss as their closure coincides with that of the set of f that we are interested in. The extremal step functions are of the form

$$f(x) = \begin{cases} 1, & 0 \leq x < c_1, \\ a, & c_1 \leq x < c_2, \\ 0, & c_2 \leq x, \end{cases} \tag{4.18}$$

where $c_1 \geq 0$, $c_1 \leq c_2 < \frac{\pi}{2}$, $0 < a < 1$, and f must integrate out to the desired solid angle Θ. At this stage one can certainly simulate the minimal NSM problem on a computer, using the above consideration together with the equations (4.15), (4.16), (4.17). We have thus numerically verified the truth of Conjecture 3 for various choices of d. The concavity hypothesis Conjecture 4 has a direct route to numerical verification, and we have performed this verification as well.

We now outline a possible route towards an analytical proof, at least for Conjecture 3. A natural guess is that reducing the length of the $[c_1, c_2]$ interval in (4.18) and increasing a accordingly could monotonically decrease the NSM as a increases. Once translated into mathematics, we have the following conjecture (that would therefore imply Conjecture 3). Conjecture 5 given below is one of the technical inequalities alluded to earlier. The other inequality corresponds to verifying the concavity hypothesis given in Conjecture 4.

Conjecture 5. *Let $0 \leq a < b < \frac{\pi}{2}$, and let $d \geq 2$. Then*

$$\frac{\int_a^b \sin(t)^{d-2}(\sec(b)^{d+2} - \sec(t)^{d+2})dt}{\int_a^b \sin(t)^{d-2}(\sec(b)^{d} - \sec(t)^{d})dt} \geq \left(1 + \frac{2}{d}\right) \frac{\int_a^b \sin(t)^{d-2}\sec(t)^{d+2}dt}{\int_a^b \sin(t)^{d-2}\sec(t)^{d}dt} \tag{4.19}$$

Remark 12. *A reader may wonder whether one needs to use the decreasing nature of f, which was slightly less trivial to see than $0 \leq f \leq 1$. Retracing our steps with this larger convex set instead, one ends up with another inequality of similar flavor to* (4.19)*, given by*

$$\frac{\sec(b)^{d+2} - \sec(a)^{d+2}}{\sec(b)^d - \sec(a)^d} \geq \frac{d+2}{d} \frac{\int_a^b \tan(t)^{d-2} \sec(t)^4 dt}{\int_a^b \tan(t)^{d-2} \sec(t)^2 dt}. \tag{4.20}$$

Unfortunately, (4.20) *is false for $d > 3$, though it does appear to be true for $d = 2, 3$! Geometrically, this means that the NSM of an annular cone does not necessarily decrease as the annulus is brought towards the normal while keeping the solid angle fixed for $d \geq 4$. We do note that, not surprisingly,* (4.19) *is a weaker inequality than the (too strong and incorrect)* (4.20)*. The fact that* (4.19) *can be deduced from the (incorrect for $d > 3$!)* (4.20) *comes from the fact that*

$$\frac{\sec^{d+2}(b) - \sec^{d+2}(x)}{\sec^d(b) - \sec^d(x)}$$

is increasing in x.

In our attempts to prove Conjecture 5, we attempted proving a more general statement that also appears to be true numerically: Let

$$f(\alpha, \beta; a, b) \triangleq \frac{\int_a^b \sin(t)^\alpha (\sec(b)^\beta - \sec(t)^\beta) dt}{\beta \int_a^b \sin(t)^\alpha \sec(t)^\beta dt}. \tag{4.21}$$

Then the claim is that $f(\alpha, \beta; a, b)$ is monotonically increasing in β. One may also favor a $u = \sec(t)$ substitution to convert this general statement into showing monotonicity in β for $1 \leq a \leq b < \infty$ of

$$g(\alpha, \beta; a, b) \triangleq \frac{\int_a^b (u^2-1)^{\frac{\alpha-1}{2}} u^{-\alpha-1} (b^\beta - u^\beta) du}{\beta \int_a^b (u^2-1)^{\frac{\alpha-1}{2}} u^{\beta-\alpha-1} du}. \tag{4.22}$$

Finally, we do note that although our "monotonicity approach" outlined in Conjecture 5 does not prove the concavity hypothesis given in Conjecture 4, it does still *have at least one interesting consequence other than Conjecture 3, namely that*

$$\frac{\phi_d(\Theta)^{\frac{d}{2}}}{\Theta}$$

is nonincreasing (take $c_1 = 0$ in (4.18)*). The astute reader may recall this as an easy consequence of concavity (or equivalently a weaker variant thereof) that was also utilized in Newman's proof.*

4.4 Discussion and Future Work

Our original impetus for undertaking the study of lower bounds on the minimum mean squared error of vector quantizers was inspired by the following folklore

Conjecture 6. *E_8 and the Leech lattice are optimal vector quantizers in the mean squared error, high resolution limit sense (G_d) for $d = 8$ and $d = 24$ respectively.*

Given the relatively recent resolution of the sphere packing problem and associated universal optimality phenomena for $\mathbb{R}^d$, $d = 8, 24$ [120], [27], [28], we believed it worthwhile to study the main Conjecture 6. We note that there is fairly reasonable numerical evidence in support of Conjecture 6 in [2], whose main contribution is a gradient based optimization algorithm for quantizers, different from the classical Lloyd-Max [79], [81] [2]. Essentially, they perform a gradient descent on a parametrization of the space of lattices, and generalize the approach to quantizers formed by k translates of a lattice. For $d = 8, 24$, the algorithm ends up at E_8 and the Leech lattice for many choices of random starting point.

One may therefore view our contributions here as coming up with a different kind of evidence based on more rigorous considerations. For example, as Table 4.4 shows, the scaled NSM for E_8 is ≈ 0.07168, while our conjectured lattice lower bound (4.5) is ≈ 0.07102. For $d = 24$, the gap is larger, with our conjecture giving the lower bound of ≈ 0.06561, and the Leech lattice yielding a performance of ≈ 0.06577.

[2] The Lloyd-Max algorithm goes back at least to Steinhaus [109].

d	[128] l.b.	*conj.* Λ l.b. (4.5)	[31] *conj.* l.b.	u.b.	[128] u.b.
1	**.08333**	**.08333**	**.08333**	**.08333**	.50000
2	.07958	**.08019**	**.08019**	**.08019**	.15915
3	.07697	.07773	.07787	.07854	.11580
4	.07503	.07580	.07609	.07660	.09974
5	.07352	.07425	.07465	.07563	.09132
6	.07230	.07298	.07347	.07424	.08608
7	.07130	.07192	.07248	.07273*	.08248
8	.07045	.07102	.07163	.07168	.07982
9	.06973	.07026	.07090	.07110*	.07778
10	.06910	.06959	.07026	.07081	.07614
16	.06657	.06689	.06759	.06830	.07053
24	.06475	.06497	.06561	.06577	.06722
∞	**.05855**	**.05855**	**.05855**	**.05855**	**.05855**

Table 4.1: Numerical values for various bounds on the NSM (divided by the dimension d) up to a couple of decimal places. Bold face denotes rigorously known sharp values. * denotes non-lattice constructions [2]. "l.b." and "u.b." stand for "lower bound" and "upper bound" respectively.

However, there is still a fair bit of work to be done in order to make (4.5) fully rigorous. The work there has been encapsulated in the form of two Conjectures 3, 4. It is also clear that we are quite far from establishing the folklore Conjecture 6 that provides the answer for $d = 8, 24$. We believe that establishing Conjecture 6 requires new methods, and it would be pleasant if such methods could also extend to other values of d and thereby supersede our conjectured lower bound (4.5), our rigorous general lower bound 10 (with tuned parameters), and Conway and Sloane's conjectured lower bound [31]. We believe that this search for new methods may in fact be the most fruitful approach to establishing these conjectures.

Stepping back from the considerations of mean squared error and the high resolution limit, it is interesting to understand to what extent these methods can be applied to other distortion measures.

Chapter 5

Conclusion and Future Directions

> When someone failed, another has succeeded; what was unknown in one century, the next has discovered; science and the arts do not grind themselves into uniformity, but gain shape and regularity by carving and polishing repeatedly... What my own strength has not been able to uncover, I cease not from working at and trying out and, by reshaping and solidifying this new material, in moulding and heating it, I bequeathe to him who follows some facility and make it the more supple and malleable for him. The second will do the same for the third, which is why difficulty does not make me despair, nor of my own weakness...
>
> *Michel de Montaigne*, Les Essais, Livre II, Chapitre XII

Let us summarize the main contributions of this dissertation at a very high level. We began by reviewing the notion of codes and anticodes in a metric space in Chapter 2. We then proceeded to review the notions of pairwise potential energy, ground states, and universal optimality. We also reviewed the notion of noise stability popular in theoretical computer science. We then reviewed and discussed Fourier analysis on finite groups, and how one can derive certain linear programming bounds. We used these bounds to prove that certain natural Boolean functions maximize noise stability subject to an expected value constraint.

In Chapter 3, we considered the problem of maximizing the quality of reconstruction for a coded aperture imaging apparatus under a simple model. We described how this problem boils down to the question of how and to what extent can one shape the magnitudes of the Fourier coefficients on $\mathbb{Z}/n\mathbb{Z}$ subject to an l_∞ constraint in time. We thus constructed a link with the "coefficient problem" in harmonic analysis, and accordingly utilized Nazarov's solution [86] in the resolution. We also showed how one can make Nazarov's solution algorithmically effective.

In Chapter 4, we considered the problem of finding improved lower bounds on the mean squared error of vector quantizers in the so-called "high-resolution limit" where one lets the number of quantizer points per unit volume tend to infinity. We developed two approaches: one to handle lower bounds on lattice quantizers, and the other to handle lower bounds on general quantizers. The lower bound we obtain for general quantizers is rigorous, while the one for lattices rests on certain plausible and easily numerically verified conjectures 3, 4 that we are currently unable to prove.

Throughout this dissertation, we have been guided by a couple of principles and themes. In particular, we have been inspired by the research philosophy of Yuri Vladimirovich Linnik. According to [66]:

"Linnik often liked to say that when starting a new area of research one should select in it a difficult and neatly formulated problem: in trying to solve it, new problems will crop up and the problem itself will serve as a touchstone for the methods being used. This would lead step-by-step to the creation of a theory and of general methods."

In order to execute upon this program, in this dissertation we focused on topics and problems with a rich history. This has the positive effect of allowing one to focus on coming up with syntheses of methods as well as formulating "new" methods. However, this may come at a cost of not addressing the most relevant problems of our current era. Our general response to such a criticism is twofold. First, we believe it shortsighted to lay judgement upon relevance based on our current era as it is close to impossible to anticipate the future. Second, ultimately progress is achieved through new methods and ideas. Problems that are classical, difficult, and neatly formulated have been proven over time to be essentially as fruitful towards this ultimate goal as the formulation of new fields and disciplines. With this view in mind, historical anecdotes and references are collectively another principal contribution of this dissertation.

Each chapter of this dissertation closed with its own suggestions for future research, and we do not wish to repeat specific technicalities here. As such, we close with just a few, very high level, ideas:

1. One key underlying theme here was the use of Fourier analysis to attack problems of geometric character. Such a program has been pursued since the inception of Fourier analysis. However, we believe that we are still at a very early stage here and anticipate substantial advances in the future. For example, the linear programming bounds have been primarily used for coding theory questions. We have demonstrated in this dissertation that they can be used for isodiametric questions (see also [124, 49]), and also vector quantization questions. Another illustration of the interplay is provided by hypercontractivity, a topic which we do not explore in this dissertation. We look forward to a synthesized, general point of view that encompasses both the linear programming bounds and hypercontractivity.

2. Another theme here is the understanding of how "bulk" constraints in time (such as l_p) translate into "fine-grained" constraints on frequency components. For example, Nazarov's theorem refers to the individual frequency components, and not just their "bulk" characteristics that can be captured by e.g. l_q norms. These "bulk" quantities are covered by more classical theory on the $l_p \to l_q$ operator norms; see for example work on hypercontractivity. Can we obtain further understanding of the fine-grained structure of the frequency components? For example, can we usefully incorporate phase information instead of just discussing magnitudes?

3. What are the broader scientific and engineering implications of answers and the search for answers to the preceding items? We have no idea and we generally devote greater energy to the preceding items instead as they are more easily formulated. Nevertheless, we hope that the reader pleasantly surprises us!

Appendices

Appendix A

Proofs for Chapter 3

As remarked in the main text, Lemma 24 is really just a calculus exercise that offers limited insight. For example, one can reduce this to studying a single variable function of a, and finding its maximum. This may be easily done on a computer to whatever degree of precision is desired, and such a numerical study has been performed in our code:`https://github.com/gajjanag/apertures`.

Nevertheless, we give an unenlightening fully rigorous "analytical" proof below for completeness. This argument naturally did not appear in our paper [8] due these reasons as well as space constraints.

Proof of Lemma 24. First, one may compute

$$\begin{aligned} f_a'(x) &= \frac{-2ax - x^2 + a}{(a+x)^2}, \\ f_a''(x) &= -\frac{2a(a+1)}{(a+x)^3}. \end{aligned}$$

Thus f_a is a concave function since $a > 0$. Furthermore, $f_a(0) = f_a(1) = 0$, so in fact $f_a(x)$ attains its maximum precisely at the root of $f_a'(x) = 0$ lying in $[0, 1]$, namely $x^* = \sqrt{a^2 + a} - a$. Plugging in this value, we get

$$f_a(x^*) = 2a + 1 - 2\sqrt{a(a+1)}. \tag{A.1}$$

Let us first look at

$$g(a) \triangleq \frac{f_a(x^*)}{f_a\left(\frac{1}{2}\right)}.$$

Using (A.1), we have the explicit expression

$$g(a) = (4a+2)(2a+1-2\sqrt{a(a+1)}).$$

We may differentiate to show that g is certainly decreasing, since

$$g'(a) = -\frac{2(2a+1-2\sqrt{a(a+1)})^2}{\sqrt{a(a+1)}} < 0.$$

Thus, $g(a) < g(0) = 2$. This proves $M(a, \{1/2\}) \leq 2$.

For $\rho = 1/4$, we have a similar function

$$h(a) \triangleq \frac{f_a(x^*)}{f_a\left(\frac{1}{4}\right)}.$$

Explicitly,

$$h(a) = \frac{1}{3}(16a+4)(2a+1-2\sqrt{a(a+1)}).$$

Differentiating, we get

$$h'(a) = -\frac{4(-2a-1+2\sqrt{a(a+1)})(-4a-1+4\sqrt{a(a+1)})}{3\sqrt{a(a+1)}}.$$

Examining the signs of the factors, we see that h decreases on $[0, 1/8]$, and then increases. Thus on $[0, 1/8]$, $\min(g(a), h(a)) \leq \min(h(0), g(0)) = 4/3$. For $a > 1/8$, we see that $\min(g(a), h(a)) \leq g\left(1/8\right) = 5/4 < 4/3$. This proves $M(a, \{1/4, 1/2\}) \leq 4/3$.

For $\rho = 1/8$, we have a similar function

$$k(a) \triangleq \frac{f_a(x^*)}{f_a\left(\frac{1}{8}\right)}.$$

Explicitly,

$$k(a) = \frac{1}{7}(64a+8)(2a+1-2\sqrt{a(a+1)}).$$

Differentiating, we get

$$k'(a) = -\frac{8(-2a-1+2\sqrt{a(a+1)})(-8a-1+8\sqrt{a(a+1)})}{7\sqrt{a(a+1)}}.$$

Examining the signs of the factors, we see that k decreases on $[0, 1/48]$, and then increases. Thus on $[0, 1/48]$, $\min(g(a), h(a), k(a)) \leq \min(g(0), h(0), k(0)) = 8/7$. On $[1/48, 1/8]$, $\min(g(a), h(a), k(a)) \leq \min(g(1/48), h(1/48)) = 13/12 < 8/7$. On $[1/8, 3/4]$, we see that $\min(g(a), h(a), k(a)) \leq h\,(3/4) = 1.113\cdots < 8/7 = 1.14\ldots$. For $a > 3/4$, we see that $\min(g(a), h(a), k(a)) \leq g\,(3/4) = 1.043\cdots < 8/7$. This proves $M(a, \{1/8, 1/4, 1/2\}) \leq 8/7$. Note that there was nothing special about $3/4$ in the above proof. Any number in a certain interval tuned appropriately to the above argument would work. □

Bibliography

[1] J. G. Ables. Fourier transform photography: a new method for X-ray astronomy. *Publications of the Astronomical Society of Australia*, 1(4):172–173, 1968.

[2] Erik Agrell and Thomas Eriksson. Optimization of lattices for quantization. *IEEE Transactions on Information Theory*, 44(5):1814–1828, 1998.

[3] Rudolf Ahlswede. Towards a general theory of information transfer. `https://www.youtube.com/watch?v=uQZBlcSH6gs`, July 2006.

[4] Rudolf Ahlswede, Harout K. Aydinian, and Levon H. Khachatrian. On perfect codes and related concepts. *Designs, Codes and Cryptography*, 22(3):221–237, 2001.

[5] Rudolf Ahlswede and Vladimir Blinovsky. *Lectures on advances in combinatorics.* Springer, 2008.

[6] Rudolf Ahlswede and Levon H. Khachatrian. The complete intersection theorem for systems of finite sets. *European Journal of Combinatorics*, 18(2):125–136, 1997.

[7] Rudolf Ahlswede and Levon H. Khachatrian. The diametric theorem in Hamming spaces—optimal anticodes. *Advances in Applied Mathematics*, 20(4):429–449, 1998.

[8] G. Ajjanagadde, C. Thrampoulidis, A. Yedidia, and G. Wornell. Near-optimal coded apertures for imaging via Nazarov's theorem. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7690–7694, May 2019.

[9] Omer Angel, Sébastien Bubeck, Yuval Peres, and Fan Wei. Local max-cut in smoothed polynomial time. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 429–437. ACM, 2017.

[10] Tom M. Apostol. *Introduction to analytic number theory.* Springer Science & Business Media, 2013.

[11] M. Salman Asif, Ali Ayremlou, Aswin Sankaranarayanan, Ashok Veeraraghavan, and Richard G Baraniuk. Flatcam: Thin, lensless cameras using coded

aperture and computation. *IEEE Transactions on Computational Imaging*, 3(3):384–397, 2017.

[12] Konstantin Ivanovich Babenko. An inequality in the theory of Fourier integrals. *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, 25(4):531–542, 1961.

[13] Paul Balister, Béla Bollobás, Robert Morris, Julian Sahasrabudhe, and Marius Tiba. Flat Littlewood polynomials exist, 2019.

[14] Keith Ball. A lower bound for the optimal density of lattice packings. *International Mathematics Research Notices*, 1992(10):217–221, 05 1992.

[15] Keith Ball. Convex geometry and functional analysis. *Handbook of the geometry of Banach spaces*, 1:161–194, 2001.

[16] Thøger Bang. A solution of the "plank problem". *Proceedings of the American Mathematical Society*, 2(6):990–993, 1951.

[17] William Beckner. Inequalities in Fourier analysis. *Annals of Mathematics*, 102(1):159–182, 1975.

[18] Itai Benjamini, Gil Kalai, and Oded Schramm. Noise sensitivity of Boolean functions and applications to percolation. *Publications Mathématiques de l'Institut des Hautes Études Scientifiques*, 90(1):5–43, 1999.

[19] W. R. Bennett. Spectra of quantized signals. *The Bell System Technical Journal*, 27(3):446–472, July 1948.

[20] Hans F. Blichfeldt. The minimum value of quadratic forms, and the closest packing of spheres. *Mathematische Annalen*, 101(1):605–608, 1929.

[21] Holger Boche and Ezra Tampubolon. Mathematics of signal design for communication systems. In *Mathematics and Society*, pages 185–220. European Mathematical Society Publishing House, 2016.

[22] Salomon Bochner. *Vorlesungen über Fouriersche Integrale*. Akademische Verlagsgesellschaft, 1932.

[23] Christopher M. Brown. *Multiplex imaging and random arrays*. PhD thesis, University of Chicago, 1972.

[24] S. Chowla. A property of biquadratic residues. *Proc. Nat. Acad. Sci. India. Sect. A.*, 14:45–46, 1944.

[25] Adam Lloyd Cohen. Anti-pinhole imaging. *Optica Acta: International Journal of Optics*, 29(1):63–67, 1982.

[26] Henry Cohn and Abhinav Kumar. Universally optimal distribution of points on spheres. *Journal of the American Mathematical Society*, 20(1):99–148, 2007.

[27] Henry Cohn, Abhinav Kumar, Stephen D. Miller, Danylo Radchenko, and Maryna Viazovska. The sphere packing problem in dimension 24. *Annals of Mathematics*, 185(3):1017–1033, 2017.

[28] Henry Cohn, Abhinav Kumar, Stephen D. Miller, Danylo Radchenko, and Maryna Viazovska. Universal optimality of the E_8 and Leech lattices and interpolation formulas, 2019.

[29] Henry Cohn and Yufei Zhao. Energy-minimizing error-correcting codes. *IEEE Transactions on Information Theory*, 60(12):7442–7450, 2014.

[30] J. H. Conway and N. J. A. Sloane. *Sphere Packings, Lattices and Groups.* Springer New York, New York, NY, 1999.

[31] John Conway and Neil Sloane. A lower bound on the average error of vector quantizers (corresp.). *IEEE Transactions on Information Theory*, 31(1):106–109, 1985.

[32] Harold Davenport. *Multiplicative number theory*, volume 74. Springer Science & Business Media, 2013.

[33] Karel De Leeuw, Yitzhak Katznelson, and Jean-Pierre Kahane. Sur les coefficients de Fourier des fonctions continues. *CR Acad. Sci. Paris Sér. AB*, 285(16):A1001–A1003, 1977.

[34] Boris Nikolaevich Delone and Nina Nikolaevna Sandakova. Theory of stereohedra. *Trudy Matematicheskogo Instituta imeni VA Steklova*, 64:28–51, 1961. In Russian.

[35] Philippe Delsarte. An algebraic approach to the association schemes of coding theory. *Philips Res. Reports Suppls.*, 10, 1973.

[36] Philippe Delsarte and Vladimir I. Levenshtein. Association schemes and coding theory. *IEEE Transactions on Information Theory*, 44(6):2477–2504, 1998.

[37] R.H. Dicke. Scatter-hole cameras for X-rays and gamma rays. *The astrophysical journal*, 153:L101, 1968.

[38] L. E. Dickson. Cyclotomy, higher congruences, and Waring's problem. *American Journal of Mathematics*, 57(2):391–424, 1935.

[39] Nikolai Dolbillin and Masaharu Tanemura. How many facets on average can a tile have in a tiling. *Forma*, 21(3):177–196, 2006.

[40] Marco F. Duarte, Mark A. Davenport, Dharmpal Takhar, Jason N. Laska, Ting Sun, Kevin F. Kelly, and Richard G. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 25(2):83–91, 2008.

[41] Paul Erdős. Some unsolved problems. *Michigan Math. J.*, 4(3):291–300, 1957.

[42] Paul Erdős. Intersection theorems for systems of finite sets. *Quart. J. Math. Oxford Ser.(2)*, 12:313–320, 1961.

[43] L. Fejes Tóth. Sur la représentation d'une population infinie par un nombre fini d'éléments. *Acta Mathematica Academiae Scientiarum Hungarica*, 10(3):299–304, Sep 1959.

[44] Edward E. Fenimore and Thomas M. Cannon. Coded aperture imaging with uniformly redundant arrays. *Applied optics*, 17(3):337–347, 1978.

[45] Richard P. Feynman. *Feynman lectures on physics. Volume 2: Mainly electromagnetism and matter.* Reading, Ma.: Addison-Wesley, 1964, edited by Feynman, Richard P.; Leighton, Robert B.; Sands, Matthew, 1964.

[46] Yuval Filmus. The weighted complete intersection theorem. *Journal of Combinatorial Theory, Series A*, 151:84–101, 2017.

[47] G. David Forney. *Transforms and Groups*, pages 79–97. Springer US, Boston, MA, 1998.

[48] Jean-Baptiste Joseph Fourier. *Théorie analytique de la chaleur.* F. Didot, 1822.

[49] Péter Frankl and Richard M. Wilson. The Erdős-Ko-Rado theorem for vector spaces. *Journal of Combinatorial Theory, Series A*, 43(2):228–236, 1986.

[50] Fang-Wei Fu, Victor K. Wei, and Raymond W. Yeung. On the minimum average distance of binary codes: linear programming approach. *Discrete Applied Mathematics*, 111(3):263–281, 2001.

[51] Carl Friedrich Gauß. *Disquisitiones Arithmeticae.* Lipsiae In Commissis Apud Gerh. Fleischer Jux., 1801.

[52] Allen Gersho and Robert M. Gray. *Vector quantization and signal compression*, volume 159. Springer Science & Business Media, 2012.

[53] Dion Gijswijt, Alexander Schrijver, and Hajime Tanaka. New upper bounds for nonbinary codes based on the Terwilliger algebra and semidefinite programming. *Journal of Combinatorial Theory, Series A*, 113(8):1719–1731, 2006.

[54] Ambros Gleixner, Leon Eifler, Tristan Gally, Gerald Gamrath, Patrick Gemander, Robert Lion Gottwald, Gregor Hendel, Christopher Hojny, Thorsten Koch, Matthias Miltenberger, Benjamin Müller, Marc E. Pfetsch, Christian Puchert, Daniel Rehfeldt, Franziska Schlösser, Felipe Serrano, Yuji Shinano, Jan Merlin Viernickel, Stefan Vigerske, Dieter Weninger, Jonas T. Witt, and Jakob Witzig. The SCIP optimization suite 5.0. Technical Report 17-61, ZIB, Takustr. 7, 14195 Berlin, 2017.

[55] Ambros M. Gleixner, Daniel E. Steffy, and Kati Wolter. Iterative refinement for linear programming. Technical Report 15-15, ZIB, Takustr. 7, 14195 Berlin, 2015.

[56] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete Mathematics: A Foundation for Computer Science (2nd Edition)*. Addison-Wesley Professional, 2 edition, 3 1994.

[57] R. M. Gray and D. L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383, Oct 1998.

[58] Ben Green. Spectral structure of sets of integers. In *Fourier analysis and convexity*, pages 83–96. Springer, 2004.

[59] Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005.

[60] Lawrence Hueston Harper. Optimal assignments of numbers to vertices. *Journal of the Society for Industrial and Applied Mathematics*, 12(1):131–135, 1964.

[61] F. Hausdorff. Eine ausdehnung des parsevalschen satzes über fourierreihen. *Mathematische Zeitschrift*, 16:163–169, 1923.

[62] G. Herglotz. Über potenzreihen mit positivem, reelen teil im einheitskreis. *Ber. Verhandl. Sachs Akad. Wiss. Leipzig, Math.-Phys. Kl.*, 63:501–511, 1911.

[63] Otto Hölder. Über einen mittelwertsatz. *Göttinger Nacgrichten*, pages 38–47, 1889.

[64] Jerry Lee Holsinger. Digital communication over fixed time-continuous channels with memory-with special application to telephone channels. *Research Laboratory of Electronics Technical Report*, 430, 1964.

[65] J. Huang and P. Schultheiss. Block quantization of correlated Gaussian random variables. *IEEE Transactions on Communications Systems*, 11(3):289–296, 1963.

[66] I. A. Ibragimov, S. M. Lozinskii, A. V. Malyshev, V. V. Petrov, Yu. V. Prokhorov, N. A. Sapogov, and D. K. Faddeev. Yurii Vladimirovich Linnik (obituary). *Russian Mathematical Surveys*, 28(2):197–215, April 1973.

[67] R. C. Jennison. *Fourier transforms and convolutions for the experimentalist*. Pergamon Press, Oxford, 1961.

[68] Steven Johnson. *The ghost map: The story of London's most terrifying epidemic—and how it changed science, cities, and the modern world*. Penguin, 2006.

[69] Grigorii Anatol'evich Kabatyanskii and Vladimir Iosifovich Levenshtein. On bounds for packings on a sphere and in space. *Problemy Peredachi Informatsii*, 14(1):3–25, 1978.

[70] Yitzhak Katznelson. *An Introduction to Harmonic Analysis*. Cambridge Mathematical Library. Cambridge University Press, 3 edition, 2004.

[71] Daniel J. Kleitman. On a combinatorial conjecture of Erdős. *Journal of Combinatorial Theory*, 1(2):209–214, 1966.

[72] Emmanuel Kowalski. *An introduction to the representation theory of groups*, volume 155. American Mathematical Society, 2014.

[73] Mark Krein and David Milman. On extreme points of regular convex sets. *Studia Mathematica*, 9(1):133–138, 1940.

[74] Emma Lehmer. On residue difference sets. *Canad. J. Math*, 5:425–432, 1953.

[75] Anat Levin, Rob Fergus, Frédo Durand, and William T. Freeman. Image and depth from a conventional camera with a coded aperture. *ACM transactions on graphics (TOG)*, 26(3):70, 2007.

[76] John H. Lindsey. Assignment of numbers to vertices. *The American Mathematical Monthly*, 71(5):508–516, 1964.

[77] J. E. Littlewood. On Polynomials $\sum^n \pm z^m, \sum^n e^{\alpha_m i} z^m, z = e^{\theta i}$. *Journal of the London Mathematical Society*, s1-41(1):367–376, 01 1966.

[78] John Edensor Littlewood. *Some problems in real and complex analysis.* DC Heath, 1968.

[79] Stuart P. Lloyd. Least squares quantization in PCM, 1957. unpublished Bell Lab. Techn. Note, portions presented at the Institute of Mathematical Statistics Meet., Atlantic City, NJ, Sept. 1957. Also, IEEE Trans. Inform. Theory (Special Issue on Quantization), vol. IT-28, pp. 129-137, Mar. 1982.

[80] David G. Luenberger. *Optimization by vector space methods.* Decision and control. Wiley, New York, NY, 1969.

[81] Joel Max. Quantizing for minimum distortion. *IRE Transactions on Information Theory*, 6(1):7–12, 1960.

[82] Robert McEliece, Eugene Rodemich, Howard Rumsey, and Lloyd Welch. New upper bounds on the rate of a code via the Delsarte-MacWilliams inequalities. *IEEE Transactions on Information Theory*, 23(2):157–166, 1977.

[83] R. P. Millane, S. Alzaidi, and W. H. Hsiao. Scaling and power spectra of natural images. In *Proc. Image and Vision Computing New Zealand*, pages 148–153, 2003.

[84] H. Minkowski. Allgemeine Lehrsätze über die convexen Polyeder. *Nachr. Ges. Wiss. Göttingen, Math.-Phys. Kl.*, 1897:198–219, 1897.

[85] Maho Nakata, Bastiaan J. Braams, Katsuki Fujisawa, Mituhiro Fukuda, Jerome K. Percus, Makoto Yamashita, and Zhengji Zhao. Variational calculation of second-order reduced density matrices by strong N-representability conditions and an accurate semidefinite programming solver. *The Journal of Chemical Physics*, 128(16):164113, 2008.

[86] Fedor L'vovich Nazarov. The Bang solution of the coefficient problem. *Algebra i Analiz*, 9(2):272–287, 1997. English translation in St. Petersburg Math. J. 9 (1998), no. 2, 407-419.

[87] Joseph Needham. *Science and Civilisation in China: Physics and physical technology: pt. 1. Physics, with the collaboration of Wang Ling and the special co-operation of Kenneth Girdwood Robinson*, volume 4. University Press, 1954.

[88] Donald Newman. The hexagon theorem. *IEEE Transactions on information theory*, 28(2):137–139, 1982.

[89] Ryan O'Donnell. *Analysis of Boolean functions*. Cambridge University Press, 2014.

[90] B. M. Oliver, J. R. Pierce, and C. E. Shannon. The philosophy of PCM. *Proceedings of the IRE*, 36(11):1324–1331, Nov 1948.

[91] P. F. Panter and W. Dite. Quantization distortion in pulse-count modulation with nonuniform spacing of levels. *Proceedings of the IRE*, 39(1):44–48, Jan 1951.

[92] Yury Polyanskiy and Yihong Wu. Lecture Notes on Information Theory. `http://people.lids.mit.edu/yp/homepage/data/itlectures_v5.pdf`, 2018. [Online; accessed 21-September-2018].

[93] John G. Proakis and Masoud Salehi. *Digital communications*, volume 4. McGraw-Hill New York, 2001.

[94] R. A. Rankin. On the closest packing of spheres in n dimensions. *Annals of Mathematics*, 48(4):1062–1081, 1947.

[95] Ramesh Raskar, Amit Agrawal, and Jack Tumblin. Coded exposure photography: motion deblurring using fluttered shutter. *ACM Transactions on Graphics (TOG)*, 25(3):795–804, 2006.

[96] Frédéric Riesz. Sur certains systèmes singuliers d'équations intégrales. *Annales scientifiques de l'École Normale Supérieure*, 28:33–62, 1911.

[97] Alain Robert. *Introduction to the Representation Theory of Compact and Locally Compact Groups*. London Mathematical Society Lecture Note Series. Cambridge University Press, 1983.

[98] C. A. Rogers. Lattice coverings of space. *Mathematika*, 6(1):33–39, 1959.

[99] Leonhard James Rogers. An extension of a certain theorem in inequalities. *Messenger of Math.*, 17:145–150, 1888.

[100] Walter Rudin. *Principles of mathematical analysis.* McGraw-hill New York, 3 edition, 1976.

[101] R. Salem and A. Zygmund. Some properties of trigonometric series whose terms have random signs. *Acta Mathematica*, 91(1):245–301, Dec 1954.

[102] Alexander Schrijver. New code upper bounds from the Terwilliger algebra and semidefinite programming. *IEEE Transactions on Information Theory*, 51(8):2859–2866, 2005.

[103] Jean-Pierre Serre. *Linear representations of finite groups*, volume 42. Springer, 1977.

[104] Igor Shinkar. Intersecting families, independent sets and coloring of certain graph products. Master's thesis, Weizmann Institute of Science, 2009.

[105] Daniel Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 296–305. ACM, 2001.

[106] A. J. Stam. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, 2(2):101–112, 1959.

[107] Elias M. Stein and Rami Shakarchi. *Fourier analysis: an introduction*, volume 1 of *Princeton Lectures in Analysis.* Princeton University Press, 2011.

[108] William Stein. *Elementary number theory: primes, congruences, and secrets: a computational approach.* Springer Science & Business Media, 2008.

[109] Hugo Steinhaus. Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci*, IV(C1. III):801–804, 1956.

[110] Michel Talagrand. How much are increasing sets positively correlated? *Combinatorica*, 16(2):243–258, 1996.

[111] Terence Tao. The Euler-Maclaurin formula, Bernoulli numbers, the zeta function, and real-variable analytic continuation. `https://tinyurl.com/ybweghs5`. [Online; accessed 02-October-2018].

[112] Audrey Terras. *Fourier Analysis on Finite Groups and Applications.* London Mathematical Society Student Texts. Cambridge University Press, 1999.

[113] Christos Thrampoulidis, Gal Shulkind, Feihu Xu, William T. Freeman, Jeffrey H. Shapiro, Antonio Torralba, Franco N. C. Wong, and Gregory W. Wornell. Exploiting occlusion in non-line-of-sight active imaging. *IEEE Transactions on Computational Imaging*, 4(3):419–431, 2018.

[114] Kit Tiyapan. *Voronoi Translated: Introduction to Voronoi Tessellation and Essays by G.L. Dirichlet and G.F. Voronoi.* God's Ayudhya's Defence, 2010.

[115] Antonio Torralba and William T. Freeman. Accidental pinhole and pinspeck cameras: Revealing the scene outside the picture. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 374–381. IEEE, 2012.

[116] László Fejes Tóth. *Lagerungen in der Ebene auf der Kugel und im Raum.* Springer Berlin Heidelberg, Berlin, Heidelberg, 1953.

[117] Stefan Kohl (`https://mathoverflow.net/users/28104/stefan-kohl`). Existence of polynomials of degree ≥ 2 which represent infinitely many prime numbers. `https://mathoverflow.net/q/208614`. [Online; accessed 02-October-2018].

[118] Jacobus Hendricus Van Lint. *Introduction to coding theory*, volume 86. Springer Science & Business Media, 2012.

[119] Ashok Veeraraghavan, Ramesh Raskar, Amit Agrawal, Ankit Mohan, and Jack Tumblin. Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. In *ACM transactions on graphics (TOG)*, volume 26, page 69. ACM, 2007.

[120] Maryna S. Viazovska. The sphere packing problem in dimension 8. *Annals of Mathematics*, 185(3):991–1015, 2017.

[121] Georges Voronoï. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. II: Recherches sur les paralléloèdres primitifs. *J. Reine Angew. Math.*, 134:198–287, 1908.

[122] Georges Voronoï. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. premier mémoire. sur quelques propriétés des formes quadratiques positives parfaites. *J. Reine Angew. Math.*, 133:97–102, 1908.

[123] Georges Voronoï. Nouvelles applications des paramètres continus à théorie des formes quadratiques. deuxième mémoire. recherches sur les paralléloèdres primitifs. Seconde partie. *J. Reine Angew. Math.*, 136:67–182, 1909.

[124] Richard M. Wilson. The exact bound in the Erdős-Ko-Rado theorem. *Combinatorica*, 4(2-3):247–257, 1984.

[125] Adam Yedidia, Christos Thrampoulidis, and Gregory Wornell. Analysis and optimization of aperture design in computational imaging. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4029–4033. IEEE, 2018.

[126] M. Young. Pinhole optics. *Applied Optics*, 10(12):2763–2767, 1971.

[127] W. H. Young. On the Determination of the Summability of a Function by Means of its Fourier Constants. *Proceedings of the London Mathematical Society*, s2-12(1):71–88, 01 1913.

[128] P. Zador. *Development and evaluation of procedures for quantizing multivariate distributions.* PhD thesis, Stanford University, 1963.

[129] P. Zador. Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Transactions on Information Theory*, 28(2):139–149, March 1982.

[130] R. Zamir and M. Feder. On lattice quantization noise. *IEEE Transactions on Information Theory*, 42(4):1152–1159, July 1996.

[131] Changyin Zhou, Stephen Lin, and Shree Nayar. Coded aperture pairs for depth from defocus and defocus deblurring. *International journal of computer vision*, 93(1):53–72, 2011.