SOME APPLICATIONS OF RELATIVE ENTROPY IN ADDITIVE COMBINATORICS

J. WOLF

Abstract. This survey looks at some recent applications of relative entropy in additive combinatorics. Specifically, we examine to what extent entropy-increment arguments can replace or even outperform more traditional energy-increment strategies or alternative approximation arguments based on the Hahn-Banach theorem.

Contents

1. Introduction

Entropy has a long history as a tool in combinatorics. Starting with a well-known estimate for the sum of the first few binomial coefficients, some of the classical applications include Spencer's theorem that six standard deviations suffice, which states that given n finite sets, there exists a two-colouring of the elements such that all sets have discrepancy at most $\ll n^{1/2}$; a proof of the Loomis-Whitney inequality, which gives an upper bound on the volume of an n-dimensional body in Euclidean space in terms of its $(n-1)$-dimensional projections; or Radhakrishnan's proof [33] of Bregman's theorem on the maximum permanent of a 0/1 matrix with given row sums. For a beautiful introduction to these fascinating applications, as well as an extensive annotated bibliography, see [10].

There are other more recent results in additive combinatorics in particular where the concept of entropy has played a crucial role. Notable examples include Fox's improvement [7] of the bounds in the graph removal lemma (see also [30], which appeared in the proof-reading stages of this article); Szegedy's information-theoretic approach [41] to Sidorenko's conjecture (see also the blog post [14] by Gowers); Tao's solution [45] to the Erdős discrepancy problem (see the discussion [44] on Tao's blog).

Since the above developments appear to be well captured by discussions online, we shall not cover them in any detail here. Instead we shall focus on a particular strand of recent results in additive combinatorics that could all be described as "approximation theorems" of a certain kind.

The text naturally splits into five parts. To start with we give a very brief introduction to the concept of entropy and its variants, in particular relative entropy (also known as Kullback-Leibler divergence). In Section 3, we state and prove a rather general sparse approximation theorem due to Lee [28]. In Section 4 we show how it can be used to derive a variant of Chang's theorem [3] on the dimension of the large Fourier spectrum of a set, as well as Bloom's [2] more recent and powerful refinement. We subsequently derive a quadratic decomposition theorem in Section 5, which is the only part of this survey for which we claim any originality. Finally, in Section 6, we discuss, following Vadhan and Zheng [48], how the notion of relative entropy can been used to derive an optimal version of the transference principle (also known as the dense model theorem), which has had far-reaching applications in number theory, graph theory, and theoretical computer science in recent years.

In the latter two applications, at least two alternative approaches exist in the literature. One is an iterative (and in principle) constructive approach based on an ℓ^2 (or energy) increment. The idea behind it, which goes back at least as far as Szemerédi's regularity lemma, was used by Green and Tao in proving the first quadratic decompositions, as well as the original version of their transference principle. As observed by Fox [7], it sometimes turns out to be quantitatively advantageous to aim for entropy increments instead. For a discussion of the different types of increment strategies, see also [42].

An alternative approach to quadratic decompositions was pioneered by Gowers and the author in [15], and was based on the Hahn-Banach theorem (or the duality of linear programming). Gowers also used it to give an alternative proof of the transference principle [13]. While versatile and relatively clean, it suffers the disadvantage of being non-constructive. In the final section we shall see how the notion of relative entropy can

be used to give a constructive proof of the Hahn-Banach theorem (in the form of a Min-Max statement from game theory).

The purpose of this article is thus to advertise relative entropy as a powerful tool in arithmetic combinatorics, and to present the above sphere of results in a unified framework.

Acknowledgements. The author would like to thank Thomas Bloom, James Lee, Luka Rimanic, Tom Sanders and Madhur Tulsiani for helpful conversations on various aspects of this text. She is greatly indebted to James Lee for insightful comments on an earlier draft of this manuscript, and to the anonymous referee for numerous suggestions which improved the presentation.

2. A very brief introduction to entropy

Throughout we use the expectation operator $\mathbb{E}_{x\in X}$ for any finite set X to denote the normalised (finite) sum over all elements $x \in X$, that is,

$$\mathbb{E}_{x\in X} g(x) := \frac{1}{|X|}\sum_{x\in X} g(x).$$

For a function $f : X \to \mathbb{C}$, define its L^p norm as $\|f\|_p := (\mathbb{E}_{x\in X}|f(x)|^p)^{1/p}$. We also have an inner product $\langle f, g\rangle := \mathbb{E}_{x\in X} f(x)\overline{g(x)}$, which gives rise to a Hilbert space of real/complex-valued functions on X. We denote the set of measures on X by

$$\Delta_X := \{f : X \to [0,\infty) : \|f\|_1 = 1\}.$$

This family Δ_X contains, in particular, the so-called characteristic measure μ_A for any subset $A \subseteq X$, which is defined for each $x \in X$ by $\mu_A(x) := \alpha^{-1}1_A(x)$, where 1_A is the characteristic function of the set A, and $\alpha := |A|/|X|$ is its density. (When $A = \emptyset$, then μ_A is identically zero.) The *entropy* of a distribution measures its information content, or the expected amount of surprise upon observing an event in a probability space. In fact, the shape of the entropy function can be derived from a set of natural conditions (see for example [36], Chapter 9).

Definition 2.1 (Entropy)**.** *For $f \in \Delta_X$, define the* entropy *of f to be*

$$\mathrm{Ent}(f) := \mathbb{E}_{x\in X} f(x)\log f(x).$$

It is not difficult to check that for $f = \mu_A$, the characteristic measure of a subset $A \subseteq X$, we have $\mathrm{Ent}(f) = \log(\alpha^{-1})$. We may also connect this definition to the notion of *Shannon entropy* $H(Y)$ of a random variable Y taking values in X by setting $f(x) = |X|\mathbb{P}[Y = x]$

and observing that

$$H(Y) := \sum_{x \in X} \mathbb{P}[Y = x] \log \frac{1}{\mathbb{P}[Y = x]} = \log |X| - \mathrm{Ent}(f).$$

By Jensen's inequality, $H(Y) \leq \log |X|$, and thus $\mathrm{Ent}(f) \geq 0$. In fact, $\mathrm{Ent}(f)$ can be viewed as the entropy of f relative to the uniform distribution. Relative entropy is also known as Kullback-Leibler divergence [27].

Definition 2.2 (Relative entropy)**.** *For $f, g \in \Delta_X$, we define the* Kullback-Leibler (KL) divergence *or* relative entropy *from f to g by*

$$D_{KL}(f||g) := \mathbb{E}_{x \in X} f(x) \log \frac{f(x)}{g(x)}$$

whenever $\mathrm{supp}(f) \subseteq \mathrm{supp}(g)$.

It is easy to see that $D_{KL}(f||1) = \mathrm{Ent}(f)$. Again, taking $f(x) = |X|\mathbb{P}[Y = x]$, we can recover a perhaps more familiar formulation for discrete distributions Y and Z taking values in a finite range X, namely

$$D_{KL}(Y||Z) := \sum_{x \in X} \mathbb{P}[Y = x] \log \frac{\mathbb{P}[Y = x]}{\mathbb{P}[Z = x]},$$

which by convention is $+\infty$ if $\mathrm{supp}(Y) \not\subseteq \mathrm{supp}(Z)$.

It is clear that KL divergence is not in general symmetric and thus not a metric (it does not satisfy the triangle inequality either), but it does satisfy non-negativity, and equals zero only if the distributions are identical.

Kullback-Leibler divergence is a measure of the information gained when one revises one's beliefs from the prior probability distribution Z to the posterior probability distribution Y. In other words, it is the amount of information lost when Z is used to approximate Y.

The notion of relative entropy is related to the total variation $\delta(Y, Z)$ between two random variables Y, Z via *Pinsker's inequality*

$$\delta(Y, Z) \leq \sqrt{\frac{1}{2} D_{KL}(Y||Z)}.$$

We shall not explicitly make use of this fact, or indeed more advanced entropy concepts such as conditional entropy, in what follows, but since

$$\mathbb{P}[Z \in S] - \delta(Y, Z) \leq \mathbb{P}[Y \in S] \leq \mathbb{P}[Z \in S] + \delta(Y, Z),$$

it supports the intuition that if $D_{KL}(Y||Z)$ is small and it is rare for Z to lie in some set S, then it is also rare for Y to lie in S.

One reason that entropy increments can perform quantitatively better than energy increments in some applications is that every measure is at most $\log|X|$ from the uniform measure when measured in terms of KL divergence, whereas it can be as far as $|X|^{1/2}$ in the ℓ^2-distance.

3. A sparse approximation theorem

Throughout, $\mathcal{F} \subseteq L^2(X)$ will be a collection of functions $\phi : X \to \mathbb{R}$ satisfying $\|\phi\|_\infty \leq 1$. We define the semi-norm $\|\cdot\|_{\mathcal{F}}$ by

$$\|h\|_{\mathcal{F}} := \sup_{\phi\in\mathcal{F}} |\langle h, \phi\rangle|$$

for any $h \in \Delta_X$.

Definition 3.1 (Generalised Riesz products)**.** *Let $d \geq 1$ be an integer. We say a function $R \in L^2(X)$ is a* Riesz $\mathcal{F}$-product of degree at most d *if*

$$R(x) = \prod_{i=1}^{d}(1 + \epsilon_i\phi_i(x))$$

for some $\epsilon_1, \ldots, \epsilon_d \in \{-1, 0, 1\}$ and $\phi_1, \ldots, \phi_d \in \mathcal{F}$.

Note that the ϕ_is appearing in the generalised Riesz product are not necessarily distinct, and that by definition, a Riesz product is always non-negative on X.

The main result of this section, due to Lee [28], states that any $f \in \Delta_X$ can be well approximated by a number of low-degree Riesz $\mathcal{F}$-products in the sense that the resulting error has small inner product with every member of the family $\mathcal{F}$.

Theorem 3.2 (Sparse approximation theorem)**.** *For every $0 < \epsilon < e^{-3}$ and $f \in \Delta_X$, there is a $g \in \Delta_X$ such that $\|f - g\|_{\mathcal{F}} \leq \epsilon$ and there is a subset $\mathcal{F}' \subseteq \mathcal{F}$ of size $|\mathcal{F}'| \leq 9\epsilon^{-2}\cdot\mathrm{Ent}(f)$ such that g is a non-negative linear combination of Riesz $\mathcal{F}'$-products of degree at most*

$$d \leq 18\epsilon^{-1}\cdot\mathrm{Ent}(f) + O\left(\frac{\log\epsilon^{-1}}{\log\log\epsilon^{-1}}\right).$$

Proof. For some $T > 0$ we shall define a family of functions

$$\{g_t : t \in [0, T]\} \subseteq \Delta_X$$

by setting $g_0 := 1$ and

$$g_t := \frac{\exp(\int_0^t \phi_s ds)}{\mathbb{E}\exp(\int_0^t \phi_s ds)}$$

for $t \in [0,T]$. The maps $s \mapsto \phi_s$ shall be defined to be piecewise constant on a finite sequence of intervals by the following procedure.

Procedure 3.3. *Having defined the maps $s \mapsto \phi_s$ on intervals $[0,t_1),[t_1,t_2),\dots,[t_{i-1},t_i)$ for some $i \in \mathbb{N}$ with $0 < t_1 < t_2 < \cdots < t_i$, we define t_{i+1} and ϕ_s on $[t_i,t_{i+1})$, as follows.*

If there exists $\phi \in \mathcal{F}$ such that

$$|\langle g_{t_i},\phi\rangle - \langle f,\phi\rangle| > 2\epsilon/3, \tag{3.1}$$

set

$$t_{i+1} := \inf\{t \geq t_i : |\langle g_t,\phi\rangle - \langle f,\phi\rangle| \leq \epsilon/3\},$$

and

$$\phi_s := \mathrm{sign}(\langle f - g_{t_i},\phi\rangle)\phi$$

for all $s \in [t_i,t_{i+1})$.

If there is no such ϕ satisfying (3.1) at time t_i, set $T := t_i$ and $I := i$.

While this appears to be pulled out of a hat, neither the form of g_t nor the procedure for determining the maps ϕ ought to be surprising. Indeed, the aim here is to minimise $\mathrm{Ent}(g)$ over all $g \in \Delta_X$ for which $\|f-g\|_{\mathcal{F}} \leq \eta$. The Lagrangian for this problem is $\mathcal{L}(g,\lambda) = \mathrm{Ent}(g) - \sum_{\phi\in\mathcal{F}} \lambda_\phi(\langle f-g,\phi\rangle - \eta)$, with solution $\nabla_g \mathrm{Ent}(g) = -\sum_{\phi\in\mathcal{F}} \lambda_\phi\phi$ and hence $g = \exp(\sum_{\phi\in\mathcal{F}} \lambda_\phi\phi)$ (which also explains the name *gradient descent* for this general approach to solving optimisation problems of this kind). Moreover, in order to minimise the correlation of the error term $f-g$ with elements of $\mathcal{F}$, g must absorb precisely those ϕ for which the correlation is large. The latter feature is not dissimilar to other iterative approaches, in particular the widely used ℓ^2 energy increment argument.

Interpreting the Kullback-Leibler divergence $D_{KL}(f||g_t)$ as the amount of information lost when g_t is used to approximate f, we can expect said divergence to decrease over the course of the argument, starting from $\mathrm{Ent}(f)$, which is the amount of information lost when we approximate f by the uniform distribution. As a result, we have the following bounds on the stopping time T and the number of intervals I.

Claim 3.4. *The procedure above stops at time $T \leq 3\epsilon^{-1}\cdot\mathrm{Ent}(f)$.*

PROOF OF CLAIM 3.4: For $t \in [0,T)$, we compute

$$\frac{d}{dt}D_{KL}(f||g_t) = \frac{d}{dt}\mathbb{E}[f\log\frac{f}{g_t}] = -\mathbb{E}(f/g_t)\frac{d}{dt}g_t,$$

and evaluating the derivative of g_t with respect to t as $g_t(\phi_t - \langle g_t, \phi_t\rangle)$ we obtain, after some rearranging, that

$$\frac{d}{dt}D_{KL}(f||g_t) = \langle \phi_t, g_t - f\rangle.$$

By the set-up, for $t \in [t_i, t_{i+1})$, $\phi_t = \text{sign}(\langle f - g_{t_i}, \phi\rangle)\phi$ for some ϕ such that $|\langle f - g_t, \phi\rangle| \geq \epsilon/3$ for all $t \in [t_i, t_{i+1})$, so that

$$\frac{d}{dt}D_{KL}(f||g_t) = \langle \phi_t, g_t - f\rangle = -\text{sign}(\langle g_t - f, \phi\rangle)\langle \phi, g_t - f\rangle \leq -\epsilon/3.$$

Now since $D_{KL}(f||g_0) = \text{Ent}(f)$ and $D_{KL}(f||g_t) \geq 0$ for all t, by the Mean-Value theorem the procedure must terminate in time T satisfying $0 \leq \text{Ent}(f) - \epsilon/3 \cdot T$. □

Claim 3.5. *The procedure above uses $I \leq 9\epsilon^{-2} \cdot \text{Ent}(f)$ intervals.*

PROOF OF CLAIM 3.5: We shall show that $t_i - t_{i-1} \geq \epsilon/3$ for all $i \leq I$, from which it follows that the number of intervals I satisfies $I \leq T/(\epsilon/3)$, so that the required bound follows from Claim 3.4.

Fix an interval $[t_{i-1}, t_i)$ for some $i \leq I$, and for simplicity write $\phi = \phi_{t_{i-1}}$, which by definition equals ϕ_t for all $t \in [t_{i-1}, t_i)$. Then, re-using our previous calculation of the derivative of g_t, we have for $t \in [t_{i-1}, t_i)$ that

$$\frac{d}{dt}\langle \phi, g_t\rangle = -\langle \phi, g_t(\phi_t - \langle g_t, \phi_t\rangle)\rangle = -\langle \phi^2, g_t\rangle + \langle \phi, g_t\rangle^2 \geq -\|\phi\|_\infty^2\|g_t\|_1 \geq -1.$$

Since, by definition of ϕ and t_i,

$$\langle \phi, g_{t_{i-1}}\rangle - \langle \phi, g_{t_i}\rangle = \langle \phi, f - g_{t_{i-1}}\rangle - \langle \phi, f - g_{t_i}\rangle \geq 2\epsilon/3 - \epsilon/3 = \epsilon/3,$$

we obtain, again by the Mean-Value theorem, the desired conclusion that $t_i - t_{i-1} \geq \epsilon/3$. □

Note that by virtue of the construction we have $g_T \in \Delta_X$ and $\|f - g_T\|_{\mathcal{F}} \leq 2\epsilon/3$, so the function g_T would be an ideal approximant if it could be described in the appropriate form. In order to achieve this, we shall need to truncate the exponential in its definition. Let $\psi_t := \int_0^t (1 + \phi_s)ds$ and $p_d(x) := \sum_{j=0}^d \frac{x^j}{j!}$.

Claim 3.6. *There exists an integer d such that the function*

$$g := \frac{p_d(\psi_T)}{\mathbb{E}p_d(\psi_T)} \in \Delta_X$$

satisfies $\|g - g_T\|_1 \leq \epsilon/3$ *and* $\|g - g_T\|_{\mathcal{F}} \leq \epsilon/3$.

Proof of Claim 3.6: Notice that

$$g_T = \frac{\exp(\int_0^T \phi_s ds)}{\mathbb{E}\exp(\int_0^T \phi_s ds)} = \frac{\exp(\int_0^T (1+\phi_s) ds)}{\mathbb{E}\exp(\int_0^T (1+\phi_s) ds)} = \frac{\exp(\psi_T)}{\mathbb{E}\exp(\psi_T)}$$

Set

$$d := 6T + O\left(\frac{\log \epsilon^{-1}}{\log\log \epsilon^{-1}}\right)$$

so that the remainder in the degree-d Taylor expansion of $\exp(x)$ is bounded via

$$\sup_{x\in[0,2T]} \frac{|\exp(x) - p_d(x)|}{\exp(x)} \leq \frac{(2T)^{d+1}}{(d+1)!} \leq \epsilon/6.$$

Since $\|\psi_T\|_\infty \leq 2T$, it follows that

$$\|p_d(\psi_T) - \exp(\psi_T)\|_1 \leq (\epsilon/6)\mathbb{E}\exp(\psi_T) \tag{3.2}$$

and hence

$$\|g - g_T\|_1 = \left\| \frac{p_d(\psi_T)}{\mathbb{E}p_d(\psi_T)} - \frac{\exp(\psi_T)}{\mathbb{E}\exp(\psi_T)} \right\|_1 \leq \left\| \frac{p_d(\psi_T)}{\mathbb{E}\exp(\psi_T)} - \frac{\exp(\psi_T)}{\mathbb{E}\exp(\psi_T)} \right\|_1 + \left\| \frac{p_d(\psi_T)}{\mathbb{E}p_d(\psi_T)} - \frac{p_d(\psi_T)}{\mathbb{E}\exp(\psi_T)} \right\|_1.$$

The first term is bounded above by $\epsilon/6$ by (3.2), and the second term can be written as

$$\mathbb{E}p_d(\psi_T) \left| \frac{1}{\mathbb{E}p_d(\psi_T)} - \frac{1}{\mathbb{E}\exp(\psi_T)} \right| = \left| 1 - \frac{\mathbb{E}p_d(\psi_T)}{\mathbb{E}\exp(\psi_T)} \right|,$$

Again, it follows from (3.2) that $\mathbb{E}p_d(\psi_T) \geq (1-\epsilon/6)\mathbb{E}\exp(\psi_T)$, so that the second term is also bounded by $\epsilon/6$, and thus $\|g - g_T\|_1 \leq \epsilon/3$. Since for all $\phi \in \mathcal{F}$ we have $\|\phi\|_\infty \leq 1$, we also have $\|g - g_T\|_{\mathcal{F}} \leq \|\phi\|_\infty \|g - g_T\|_1 \leq \epsilon/3$. □

It follows from the construction of g_T and Claim 3.6 that $\|f - g\|_{\mathcal{F}} \leq \|f - g_T\|_{\mathcal{F}} + \|g_T - g\|_{\mathcal{F}} \leq 2\epsilon/3 + \epsilon/3 = \epsilon$. It remains to show that g is indeed of the form claimed in Theorem 3.2.

Let $\mathcal{F}'$ be the family of functionals ϕ_s encountered in the course of the above argument, i.e. $\mathcal{F}' := \{\phi_s : s \in [0,T)\} \subseteq \mathcal{F}$. By Claim 3.5, $|\mathcal{F}'| = I \leq 9\epsilon^{-2}\mathrm{Ent}(f)$. Moreover, ψ_T is by definition of the form $\sum_{\phi\in\mathcal{F}'} \lambda_\phi(1+\phi)$ for some non-negative coefficients λ_ϕ. It follows that $p_d(\psi_T)$, and hence g, is a non-negative linear combination of Riesz $\mathcal{F}'$-products of degree at most d, as claimed. □

4. The structure of the large Fourier spectrum

In this section we shall describe a first application of Theorem 3.2, also due to Lee [28]. Let G be a finite abelian group. The *Fourier transform* $\widehat{f} : \widehat{G} \to \mathbb{C}$ of a function $f : G \to \mathbb{C}$ is then defined, for each character $\gamma \in \widehat{G}$, by the formula

$$\widehat{f}(\gamma) := \mathbb{E}_{x \in G} f(x)\gamma(x).$$

We shall sometimes abuse notation and treat characters additively. The *inversion formula* states that with the above definition of the Fourier transform, we can recover f from its Fourier coefficients via the sum

$$f(x) = \sum_{\gamma \in \widehat{G}} \widehat{f}(\gamma)\gamma(x). \tag{4.1}$$

while Plancherel's identity asserts that

$$\langle f, g\rangle = \langle \widehat{f}, \widehat{g}\rangle,$$

where

$$\langle f, g\rangle := \mathbb{E}_x f(x)\overline{g(x)} \text{ and } \langle \widehat{f}, \widehat{g}\rangle := \sum_{\gamma} \widehat{f}(\gamma)\overline{\widehat{g}(\gamma)}. \tag{4.2}$$

We shall refer to (4.2) as *Parseval's identity* whenever f and g are equal. Note that inner products on physical space are normalised, while those on frequency space are not, and the same convention applies to the $L^p(G)$ and $\ell^p(\widehat{G})$ norms, respectively.

Definition 4.1. *Let $\rho > 0$, and let $f : G \to \mathbb{C}$. Define the* ρ-large spectrum *of f to be the set*

$$\mathrm{Spec}_\rho(f) := \{\gamma \in \widehat{G} : |\widehat{f}(\gamma)| \geq \rho\|f\|_1\}.$$

We shall also write

$$\mathrm{Spec}_{>0}(f) := \{\gamma \in \widehat{G} : |\widehat{f}(\gamma)| > 0\}.$$

It is not difficult to see that the ρ-large spectrum of a bounded function cannot be too large. Indeed, by Parseval's identity we have

$$\|f\|_2^2 = \|\widehat{f}\|_2^2 \geq \sum_{\gamma \in \mathrm{Spec}_\rho(f)} |\widehat{f}(\gamma)|^2 \geq \rho^2\|f\|_1^2|\mathrm{Spec}_\rho(f)|,$$

and hence $|\mathrm{Spec}_\rho(f)| \leq \rho^{-2}\|f\|_2^2/\|f\|_1^2$. In particular, when $f = \mu_A$ is the characteristic measure of a subset $A \subseteq G$ of density $\alpha := |A|/|G|$, then $|\mathrm{Spec}_\rho(\mu_A)| \leq \rho^{-2}\alpha^{-1}$.

In general, this bound is best possible as can be seen by taking A to be a subspace of $\mathbb{F}_2^n$ of fixed codimension, for example. However, it is not the most efficient description of the large Fourier spectrum. In fact, we shall see that the set of large Fourier coefficients of a function is determined by even fewer frequencies, in the following sense.

Definition 4.2 (s-covered)**.** *Let s be a positive integer. A subset $\Gamma \subseteq \widehat{G}$ is said to be* s-covered *if there exists a subset $S \subseteq \widehat{G}$ of size $|S| \leq s$ such that*

$$\Gamma \subseteq \left\{ \sum_{\gamma \in S} \epsilon_\gamma \gamma : \epsilon_\gamma \in \{-1, 0, 1\} \right\}.$$

The following structural result concerning the large spectrum is due to Chang [3]. It has seen a number of important applications in additive combinatorics, for example to Roth's theorem [38], Freiman's theorem [3] and the structure of Boolean functions with small ℓ^1 norm [21], to name just a few.

Theorem 4.3 (Chang's theorem)**.** *Let $A \subseteq G$ be a subset of density α. For every $\rho > 0$,* $\mathrm{Spec}_\rho(\mu_A)$ *is s-covered with*

$$s \leq 18\rho^{-2} \cdot \log(\alpha^{-1}).$$

This constitutes a logarithmic improvement over the bound arising from Parseval's theorem.

The proof of Chang's theorem in [3, 17] proceeds via Rudin's inequality. A first proof using entropy, restricted to the case $G = \mathbb{F}_2^n$, was later given by Impagliazzo et al. [26]. In fact, it turns out that the two are not unrelated: Friedgut [8] showed very recently how to derive hypercontractivity estimates of Bonami-Beckner type, which generalise Rudin's inequality, using entropy-based arguments.[1]

Green [16] showed that the bound in Chang's theorem as stated is tight up to a constant. However, as we shall see, it is possible to refine Theorem 4.3 in at least two ways which turn out to be advantageous in applications. First, it is possible to relax the notion of covering somewhat; second, we may require only a large subset of the spectrum to be covered.

To examine the first strengthening we need the following (non-standard) definition.

Definition 4.4 ((s,d)-covered)**.** *Let s, d be positive integers. We say a subset $\Gamma \subseteq \widehat{G}$ is* (s,d)-covered *if there exists a subset $S \subseteq \widehat{G}$ of size $|S| \leq s$ such that*

$$\Gamma \subseteq \left\{ \sum_{\gamma \in S} \epsilon_\gamma \gamma : \epsilon_\gamma \in \mathbb{Z}, \sum_{\gamma \in S} |\epsilon_\gamma| \leq d \right\}.$$

[1] A comprehensive historical account of hypercontractivity estimates can be found in [31].

Clearly if $\Gamma \subseteq \widehat{G}$ is s-covered for some $s \in \mathbb{N}$, then it is (s,s)-covered in the sense of Definition 4.4. Also, note that in the popular model setting of $G = \mathbb{F}_2^n$, the two notions of covering introduced above coincide.

In [40] Shekredov proved the following refinement of Chang's theorem.

Theorem 4.5 (Shkredov's theorem)**.** *Let $A \subseteq G$ be a subset of density α. For every $\rho > 0$, $\mathrm{Spec}_\rho(\mu_A)$ is (s,d)-covered with*

$$s \leq 2^{30}\rho^{-2}\log(\alpha^{-1}) \quad \textit{and} \quad d \leq 8\log(\alpha^{-1}).$$

While the notion of covering is in some sense weaker than that in Chang's original theorem, it turns out that Shkredov's result performs better in certain applications. For example, one obtains a quantitative improvement in Bogolyubov's lemma, which states that for $A \subseteq G = \mathbb{Z}/p\mathbb{Z}$ with p a prime, the iterated sum set $2A-2A := \{a_1+a_2-a_3-a_4 : a_1,a_2,a_3,a_4 \in A\}$ contains a Bohr set

$$B(K,\delta) := \{x \in G : \|xt/p\| \leq \delta \text{ for all } t \in K\}$$

for some set $K \subseteq \widehat{G}$ of size $|K| \leq 2^{33}\alpha^{-1}\log(\alpha^{-1})$ and width $\delta = (2^8\log(\alpha^{-1}))^{-1}$, instead of $\delta = \alpha(2^8\log(\alpha^{-1}))^{-1}$ by an application of Chang's original result.

Following Lee [28] with some small modifications, we shall deduce a slightly weaker version of Theorem 4.5 from Theorem 3.2. Indeed, we were unable to determine at the time of writing whether the full strength of Shkredov's result can be recovered in this way.

Let $\mathcal{F} = \{\Re\gamma, \Im\gamma : \gamma \in \widehat{G}\}$, and note that this is indeed a valid choice for our family of test functions. We first prove an easy lemma which says that the non-zero spectrum of generalised Riesz $\mathcal{F}'$-products for $\mathcal{F}' \subseteq \mathcal{F}$ is efficiently covered.

Lemma 4.6. *Let $\mathcal{F}' \subseteq \mathcal{F} = \{\Re\gamma, \Im\gamma : \gamma \in \widehat{G}\}$, and let $\delta > 0$. Let $R : G \to \mathbb{R}$ be a Riesz $\mathcal{F}'$-product of degree at most d, and let $r : G \to \mathbb{R}$ be any non-negative linear combination of Riesz $\mathcal{F}'$-products of degree at most d such that $r \in \Delta_G$. Then $\mathrm{Spec}_\delta(R)$ is d-covered, and $\mathrm{Spec}_\delta(r)$ is $(2|\mathcal{F}'|,d)$-covered.*

Proof. Consider a Riesz $\mathcal{F}'$-product $R = \prod_{i=1}^d(1+\epsilon_i\phi_i)$ of degree at most d, with $\phi_1,\dots,\phi_d \in \mathcal{F}'$ (not necessarily distinct) and $\epsilon_1,\dots,\epsilon_d \in \{-1,0,1\}$. This means that for each $j = 1,\dots,d$, ϕ_j is of the form $\Re\gamma_j = \frac{1}{2}(\gamma_j+\overline{\gamma}_j)$ or $\Im\gamma_j = \frac{1}{2}(\gamma_j-\overline{\gamma}_j)$ for some character $\gamma_j \in \widehat{G}$. In particular, whenever $\widehat{R}(\gamma) \neq 0$, by uniqueness of the Fourier expansion γ is a product of at most d elements from the multiset $S' := \{\gamma_1,\dots,\gamma_d,\overline{\gamma}_1,\dots,\overline{\gamma}_d\} \subseteq \widehat{G}$. This means that $\mathrm{Spec}_{>0}(R)$ is $(2d,d)$-covered. Note that we can replace a character γ' occurring k times

in S' by $\{\gamma', \gamma'^2, \dots, \gamma'^k, \overline{\gamma}', \overline{\gamma}'^2, \dots, \overline{\gamma}'^k\}$ to obtain a set S of size at most δ, allowing us to conclude that in fact, $\mathrm{Spec}_{>0}(R)$ is d-covered. Since $\mathrm{Spec}_\delta(R) \subseteq \mathrm{Spec}_{>0}(R)$, the same is obviously true of $\mathrm{Spec}_\delta(R)$.

Suppose now that $r = \sum_i c_i R_i$ is a non-negative linear combination of Riesz $\mathcal{F}'$-products of degree at most d such that $r \in \Delta_G$. Clearly

$$1 = \|r\|_1 = \sum_i c_i \|R_i\|_1$$

and if $\gamma \in \mathrm{Spec}_\delta(r)$, then

$$|\sum_i c_i \widehat{R}_i(\gamma)| = |\widehat{r}(\gamma)| \geq \delta \|r\|_1 = \delta \sum_i c_i \|R_i\|_1,$$

It follows by the triangle inequality and averaging that there exists at least one i such that

$$|\widehat{R}_i(\gamma)| \geq \delta \|R_i\|_1,$$

i.e. $\gamma \in \mathrm{Spec}_\delta(R_i)$, and thus the first part of the argument implies that $\mathrm{Spec}_\delta(r)$ is $(2|\mathcal{F}'|, d)$-covered. □

Now Theorem 3.2 allows us to approximate any function by a small number of generalised Riesz products, implying that the spectrum of such a function can also be efficiently covered.

Proof of (a weaker version of) Theorem 4.5 using Theorem 3.2: Let $f = \mu_A$ be the characteristic measure of $A \subseteq G$, whose entropy $\mathrm{Ent}(f)$ equals $\log(\alpha^{-1})$. Set $\epsilon = \rho/4$ in Theorem 3.2 to obtain a function $g \in \Delta_G$ such that $\|f - g\|_{\mathcal{F}} \leq \rho/4$ and a subset $\mathcal{F}' \subseteq \mathcal{F} = \{\Re\gamma, \Im\gamma : \gamma \in \widehat{G}\}$ of size $|\mathcal{F}'| \leq 144\rho^{-2} \cdot \mathrm{Ent}(f)$ such that g is a non-negative linear combination of Riesz $\mathcal{F}'$-products of degree at most $d \leq 72\rho^{-1}\mathrm{Ent}(f) + O(\log \rho^{-1} / \log\log \rho^{-1})$. Since $\|f - g\|_{\mathcal{F}} \leq \rho/4$, we have $\mathrm{Spec}_\rho(f) \subseteq \mathrm{Spec}_{\rho/2}(g)$. By Lemma 4.6, $\mathrm{Spec}_{\rho/2}(g)$ is $(2|\mathcal{F}'|, d)$-covered, and hence so is $\mathrm{Spec}_\rho(f)$. □

This statement implies a version of Bogolyubov's lemma in which the width of the Bohr set is $\delta = \alpha^{1/2}(2^8 \log(\alpha^{-1}))^{-1}$, a slight gain over Chang's original result.

In many applications it is not necessary to be able to cover the entire spectrum–obtaining control over a large subset often suffices. This observation was formalised and exploited by Bloom [2], who used his refinement of Chang's theorem (Theorem 4.7 below) to derive the best known bounds in Roth's theorem to date.

Theorem 4.7 (Bloom's theorem)**.** *For every $0 < \rho < e^{-3}$ and $f \in \Delta_G$, there exists a subset $S \subseteq \mathrm{Spec}_\rho(f)$ of size*

$$|S| \geq \frac{\rho}{2}|\mathrm{Spec}_\rho(f)|$$

which is d-covered for

$$d \leq 36\sqrt{2}\rho^{-1} \cdot \mathrm{Ent}(f) + O\left(\frac{\log \rho^{-1}}{\log\log \rho^{-1}}\right).$$

Proof. We set $\epsilon = \rho/(2\sqrt{2})$ in Theorem 3.2 to obtain a function $g \in \Delta_G$ such that $\|f - g\|_{\mathcal{F}} \leq \rho/(2\sqrt{2})$, and a subset $\mathcal{F}' \subseteq \mathcal{F}$ of size $|\mathcal{F}'| \leq 72\rho^{-2} \cdot \mathrm{Ent}(f)$ such that

$$g = \sum_{i=1}^{\ell} c_i R_i,$$

where ℓ is a positive integer, the coefficients c_i are positive reals and each R_i is a Riesz $\mathcal{F}'$-product of degree at most

$$d \leq 36\sqrt{2}\rho^{-1} \cdot \mathrm{Ent}(f) + O\left(\frac{\log \rho^{-1}}{\log\log \rho^{-1}}\right).$$

Note that by Lemma 4.6, the non-zero spectrum of every Riesz product R_i is d-covered. Thus if we could show that for some i, $\mathrm{Spec}_{>0}(R_i)$ is also a reasonably large proportion of $\mathrm{Spec}_\rho(f)$, we would be done.

We achieve this by choosing an element at random from $\{1, 2, \ldots, \ell\}$, where we choose j with probability $c_j \mathbb{E}R_j$. Note that since $g \in \Delta_G$, we have that the sum of the probabilities equals $\sum_{i=1}^{\ell} c_i \mathbb{E}R_i = \mathbb{E}\sum_{i=1}^{\ell} c_i R_i = \mathbb{E}g = 1$.

Fix $\gamma \in \mathrm{Spec}_\rho(f) \subseteq \mathrm{Spec}_{\rho/2}(g)$, where the latter inclusion holds since $\|f - g\|_{\mathcal{F}} \leq \rho/(2\sqrt{2})$. Now

$$\mathbb{E}_j|\langle \gamma, R_j/(\mathbb{E}R_j)\rangle| = \sum_{i=1}^{\ell} c_i \mathbb{E}R_i |\langle \gamma, R_i/(\mathbb{E}R_i)\rangle| \geq |\langle \gamma, \sum_{i=1}^{\ell} c_i R_i\rangle| = |\langle \gamma, g\rangle| \geq \rho/2,$$

where the first expectation is a probabilistic one with respect to the random choice of j from $\{1, 2, \ldots, \ell\}$. It follows from the preceding line that for any fixed $\gamma \in \mathrm{Spec}_\rho(f)$

$$\mathbb{P}_j[\widehat{R}_j(\gamma) \neq 0] = \mathbb{P}_j[|\langle \gamma, R_j/(\mathbb{E}R_j)\rangle| > 0] \geq \rho/2,$$

i.e. that any $\gamma \in \mathrm{Spec}_\rho(f)$ lies, with probability at least $\rho/2$, in $\mathrm{Spec}_{>0}(R_i)$ for a randomly chosen i, whence $\mathbb{E}_j|\mathrm{Spec}_{>0}(R_j)| \geq (\rho/2)|\mathrm{Spec}_\rho(f)|$. We conclude that there exists a choice of j for which the desired inequality $|\mathrm{Spec}_{>0}(R_j)| \geq (\rho/2)|\mathrm{Spec}_\rho(f)|$ holds. We now set $S := \mathrm{Spec}_{>0}(R_i) \subseteq \mathrm{Spec}_\rho(f)$ to obtain the desired conclusion. □

5. Quadratic decompositions

The usefulness of the Fourier transform lies in the fact that we can decompose any bounded function on G into a weighted sum of characters, which we shall often refer to as *linear phase functions* in the sequel. Indeed, since the characters form an orthonormal basis for the space of functions $G \to \mathbb{C}$, such a decomposition is unique, and gives rise to the inversion formula (4.1). Those phases with small coefficients can often be neglected in applications, but not always.

Starting with the ground-breaking work of Gowers in [11, 12], the idea that bounded functions are well approximated by their large Fourier coefficients has been successfully generalised and led to the development of so-called *higher-order Fourier analysis* over the past decade. A higher-order analogue of the Fourier inversion formula decomposes any bounded function as a weighted sum of higher-degree *polynomial phase functions*, plus a pseudorandom error term that is negligible in a wider range of applications.

The notion of pseudorandomness that will be appropriate here is due to Gowers [11, 12].

Definition 5.1 (U^k norm)**.** *Let $k \geq 2$ be an integer. The U^k* norm *of a function $f : G \to \mathbb{C}$ is defined by the formula*

$$\|f\|_{U^k}^{2^k} := \mathbb{E}_{x,h_1,\dots,h_k \in G} \Delta_{h_1} \Delta_{h_2} \dots \Delta_{h_k} f(x),$$

where $\Delta_h f(x) := f(x)\overline{f(x+h)}$ is to be thought of as a discrete phase derivative.

It is straightforward to verify that

$$\|f\|_{U^2} = \|\widehat{f}\|_4,$$

and hence that as a measure of pseudorandomness the U^2 norm and the size of the Fourier coefficients of f are, at least in a qualitative sense, equivalent. It is also not too difficult to see, but certainly not obvious, that $\|\cdot\|_{U^k}$ is indeed a norm, and that these norms are nested in the sense that for any integer $k \geq 2$,

$$\|f\|_{U^2} \leq \|f\|_{U^3} \leq \dots \leq \|f\|_{U^k} \leq \|f\|_\infty.$$

We leave the details to the keen reader, who may wish to refer to the book [43] or the lecture notes [20].

The uniformity norms defined above are useful because of the following proposition, again due to Gowers, which states that error terms which are small in U^{k+1} are negligible when it comes to counting linear patterns of complexity k. A useful example of a linear pattern of complexity k which the reader may wish to bear in mind is an arithmetic progression

of length $k+2$, but we shall not define the notion of "complexity" formally here as it is somewhat involved [24].

Proposition 5.2. *Let $\mathcal{L}$ be a linear pattern of complexity at most k, and let $f : G \to \mathbb{C}$ be a function satisfying $\|f\|_\infty \leq 1$. Then*

$$|\mathbb{E}_{x_1,\dots,x_d} \prod_{i=1}^{m} f(L_i(x_1,\dots,x_d))| \leq \|f\|_{U^{k+1}}.$$

This proposition implies that the count of 3-term arithmetic progressions (a linear pattern of complexity $k=1$) in a subset $A \subseteq G$ is controlled by the U^2 norm of the balanced function of A, and hence by the size of its Fourier coefficients. In what follows, we shall restrict our attention to the case $k=2$, and say that a function is *quadratically uniform* if it is small in the U^3 norm. In particular, Proposition 5.2 implies that any subset $A \subseteq G$ of density α in G contains the expected number of 4-term arithmetic progressions, namely $\alpha^4|G|^2$, whenever its balanced function is quadratically uniform.

When the balanced function fails to be quadratically uniform, i.e. when its U^3 norm is non-negligible, we can invoke Theorem 5.3 below, which says that such a function must correlate with a *quadratic phase function*, and hence the set A must be somewhat quadratically structured. For simplicity of the exposition we concentrate on the so-called finite-field model setting here, see [18, 49]. Theorem 5.3 as stated is due to Samorodnitsky [37] ($p=2$) and Green and Tao [22] ($p>2$), but the sophisticated ideas on which the proofs build already formed the cornerstone of Gowers's proof of Szemerédi's theorem [11].

Theorem 5.3 (U^3 inverse theorem)**.** *Let $\delta > 0$, and let $f : \mathbb{F}_p^n \to \mathbb{C}$ be a function satisfying $\|f\|_\infty \leq 1$ and $\|f\|_{U^3} \geq \delta$. Then there exists a quadratic form q over $\mathbb{F}_p^n$ such that*

$$|\mathbb{E}_x f(x)\omega^{q(x)}| \geq c(\delta),$$

where $\omega := \exp(2\pi i/p)$, for some constant $c(\delta)$ going to 0 as δ tends to 0.

So far we have expressed the dichotomy between quadratic uniformity and quadratic structure as two separate statements about the U^3 norm of a function. It is often convenient to combine these into one, by decomposing the function into a quadratically uniform and a quadratically structure part, as alluded to in the introductory paragraph of this section. Indeed, in the past decade a significant amount of work has been done towards justifying the labels *quadratic* and *higher-order Fourier analysis* by obtaining quadratic

and higher-order decompositions that (at least weakly) possess many of the useful properties that the traditional Fourier decomposition exhibits. One of the first explicit quadratic decomposition theorems was provided in [20].

Theorem 5.4 (Green-Tao decomposition)**.** *Let $f : \mathbb{F}_p^n \to [-1,1]$ be a function and let $\delta > 0$. Then there is a so-called quadratic factor (B_1, B_2) of complexity at most $O_\delta(1)$ such that*

$$f = \mathbb{E}(f|(B_1, B_2)) + g,$$

where $\|g\|_{U^3} \leq \delta$.

A *quadratic factor (B_1, B_2) of complexity $O_\delta(1)$* is a simultaneous level set of $O_\delta(1)$ linear and quadratic phases, and the projection $\mathbb{E}(f|(B_1, B_2))$ of f onto such a factor is simply f averaged over each of the level sets comprising the factor. For a precise definition, see [20]. The proof of Theorem 5.4 uses an ℓ^2 energy increment and makes crucial use of Theorem 5.3 above.

A quadratic decomposition which more closely resembles that obtained in classical Fourier analysis is due to Gowers and the author [15][2].

Theorem 5.5 (Gowers-W. decomposition)**.** *Let $f : \mathbb{F}_n^p \to \mathbb{C}$ be a function such that $\|f\|_2 \leq 1$. Then for every $\epsilon > 0$ and $\eta > 0$, there exists $M = \exp(O(\log(\eta\epsilon)^{-O(1)}))$ such that f has a decomposition of the form*

$$f = \sum_i \lambda_i \omega^{q_i} + h + \ell,$$

where the q_i are quadratic forms on $\mathbb{F}_p^n$, $\|h\|_{U^3} \leq \epsilon$, $\|\ell\|_1 \leq \eta$ and $\sum_i |\lambda_i| \leq M$.

Under the so-called Polynomial Freiman-Ruzsa Conjecture (see [18, 49]) one might expect a polynomial tradeoff between the uniformity of the function h and the complexity M of the quadratically structured part of the above decomposition. The proof of Theorem 5.5 was based on Theorem 5.3 and the finite-dimensional Hahn-Banach theorem (see also Section 6).

In [47], a different technique from machine learning, known as *boosting*[3], was used to obtain the following variant.

[2]The statement given incorporates Sanders's quantitative improvement on the Bogolyubov lemma in [39], and is thus different from the published version in [15].

[3]In fact, this technique is closely related to the gradient descent approach described in Section 3, so Theorem 5.7 should not come as a surprise.

Theorem 5.6 (Tulsiani-W. decomposition)**.** *Let $\mathcal{F}$ be a class of functions $\phi : X \to [-1,1]$ closed under negation, and let $\epsilon, \delta > 0$ and $B > 1$. Let A be an algorithm which, given oracle access to a function $f : X \to [-B,B]$ satisfying $\|f\| \geq \epsilon$, outputs, with probability at least $1-\delta$, a function $\phi \in \mathcal{F}$ such that $\langle f, \phi\rangle \geq \eta$ for some $\eta = \eta(\epsilon, B)$. Then there exists an algorithm which, given any function $f : X \to [-1,1]$, outputs with probability at least $1 - \eta^{-2}\delta$ a decomposition*

$$f = \sum_{i=1}^{k} c_i \phi_i + h + \ell$$

satisfying $k \leq \eta^{-2}$, $\|h\| \leq \epsilon$, $\|\ell\|_1 \leq (2B)^{-1}$ and $\phi_i \in \mathcal{F}$ for all $i = 1, \ldots, k$. The algorithm makes at most k calls to A.

The norm $\|\cdot\|$ can be taken to be $\|\cdot\|_{\mathcal{F}}$, but also the U^3 norm. An explicit probabilistic algorithm A was given in the paper. Note that when such an algorithm is deterministic, the resulting decomposition algorithm is deterministic. One of the advantages of Theorem 5.6 over Theorem 5.5 was that it naturally gave a bound on the number of quadratic phases involved in the decomposition, which the authors had to work much harder to obtain in [15].

As an immediate application of the sparse approximation theorem in Section 3 (Theorem 3.2) we are able to deduce yet another quadratic decomposition theorem.

Theorem 5.7. *Let $f : \mathbb{F}_p^n \to \mathbb{R}^+$ be a function such that $\|f\|_1 \leq 1$, and let $0 < \epsilon < e^{-3}$. Then there exists a function $g : \mathbb{F}_p^n \to \mathbb{R}^+$ such that $\mathbb{E}f = \mathbb{E}g$, and a set $\mathcal{Q}'$ of quadratic forms of size $|\mathcal{Q}'| \leq 9\epsilon^{-2}\mathrm{Ent}(f/\|f\|_1)$ such that f can be written as*

$$f = g + h,$$

where $g = \sum_i \lambda_i \omega^{q_i}$ with real coefficients λ_i, $\|h\|_{\mathcal{Q}} \leq \epsilon$ and each q_i is of the form

$$q_i = \sum_{q' \in Q_i'} q' - \sum_{q'' \in Q_i''} q'',$$

where $Q_i', Q_i'' \subseteq \mathcal{Q}'$ are (multi)sets whose sizes are bounded above by

$$|Q_i'| + |Q_i''| \leq 18\epsilon^{-1}\mathrm{Ent}(f/\|f\|_1) + O\left(\frac{\log(\epsilon^{-1})}{\log\log(\epsilon^{-1})}\right).$$

Proof. Let $\mathcal{F} := \{\Re\omega^{q(x)}, \Im\omega^{q(x)} : q \text{ a quadratic form on } \mathbb{F}_p^n\}$, and use Theorem 3.2 applied to $f/\|f\|_1$ to obtain a family $\mathcal{F}' \subseteq \mathcal{F}$, which gives rise to a family of quadratic forms $\mathcal{Q}'$,

and a function $g' : \mathbb{F}_p^n \to \mathbb{R}^+$ such that $g' \in \Delta_G$ and g' is a non-negative linear combination of Riesz $\mathcal{Q}'$-products. It is easy to check that $g := g'\|f\|_1$ satisfies the stated conditions. □

Theorem 5.7 could be interpreted as a quadratic Chang-type theorem as it states that the large quadratic Fourier spectrum is spanned by a small number of quadratic phases. But in some respects it is weaker than the above-cited quadratic decomposition theorems: it does not give a bound on the number of terms in the structured part of the decomposition, and the error h is only guaranteed to not have any correlation with a quadratic phase (rather than being small in U^3).

6. The transference principle

In view of the application described in the previous section, it seems natural to try and apply the entropy-increment method to derive a third type of result in arithmetic combinatorics which is generally accessible by either the ℓ^2 energy increment or the Hahn-Banach approach: the *transference principle* states, informally speaking, that any distribution which is dense with respect to a pseudorandom measure can be well approximated by a genuinely dense distribution. We phrase this principle here in such a way that it naturally fits with the content of the previous sections. As before, let $\mathcal{F}$ be a family of functions $\{\phi : X \to [0,1]\}$, and for every integer $k \geq 1$, define $\mathcal{F}^k$ to be the set of all k-fold products of elements of $\mathcal{F}$.

Theorem 6.1 (Transference principle, I)**.** *For all $\epsilon > 0$, there exists $k = \epsilon^{-O(1)}$ and $\eta = \exp(-\epsilon^{-O(1)})$ such that the following holds.*

For any family $\mathcal{F} = \{\phi : X \to [0,1]\}$ and any measure $\nu : X \to \mathbb{R}^+$ which is η-pseudorandom with respect to $\mathcal{F}^k$ in the sense that $\|\nu - 1_X\|_{\mathcal{F}^k} \leq \eta$, and any function $f : X \to \mathbb{R}^+$ with $0 \leq f \leq \nu$ and $\mathbb{E}f \leq 1$, there exists a bounded function $g : X \to [0,1]$ such that $\mathbb{E}g = \mathbb{E}f$ and $\|f - g\|_{\mathcal{F}} \leq \epsilon$.

Such a statement was first introduced by Green [19] in his work on 3-term arithmetic progressions in the primes, in the case where the family $\mathcal{F}$ consisted of linear characters and ν was a majorant of the primes, based on the set of almost-primes which exhibits all the right pseudorandomness properties. The Fourier-analytic manifestations of the transference principle are well summarised in Prendiville's recent survey [32]. Its true power, however, only came to light a few years later in Green and Tao's celebrated proof that there are arbitrarily long arithmetic progressions in the primes [23]. The principle was made explicit for the first time by Tao and Ziegler in [46], and a little later given a new

proof by Gowers [13] and independently Reingold, Trevisan, Tulsiani and Vadhan [35], who introduced the Hahn-Banach (or duality of linear programming) approach which greatly improved the quantitative dependence of the parameters involved. This improved dependence turned out to be crucial for applications in theoretical computer science, where the transference principle is now widely known as the *dense model theorem*. Its applications include but are not limited to leakage-resilient cryptography and connections to computational differential privacy (for references, see [48]). Reingold et al. [35] also investigated the broader implications of their approach: they obtained a new proof of Impagliazzo's hard core set theorem, which states that if a Boolean function is somewhat hard on average, then there must be a subset of inputs (the hard core) on which it is extremely hard, and outside of which it is easy; and a new proof of Frieze and Kannan's weak regularity lemma [9], which allows one to partition the vertex set of any graph into a small number of parts in such a way that the edge density between any two subsets of vertices in the graph is close to the value expected, based on the edge densities between the parts of the partition and the intersection sizes of the vertex subsets within these parts.

To round off this historical account, let us mention that new light was shed on the exact nature of the pseudorandomness conditions needed for the transference principle to work by Conlon, Fox and Zhao [5] (see also the expository note [51] by Zhao), and indeed their expository article [4] is the most compact and up-to-date reference at the time of writing. In fact, their work furnishes a very general regularity and counting lemma in sparse pseudorandom hypergraphs, which allows one to prove analogues of well-known combinatorial theorems such as Ramsey's theorem and Turán's theorem relative to certain sparse pseudorandom hypergraphs.

The remainder of this section is structured as follows. First, we give a proof of Theorem 6.1 following [34] which relies on the finite-dimensional Hahn-Banach theorem, presented here in the (equivalent) form of the Min-Max theorem from game theory. In the second half of this section we describe recent work of Vadhan and Zheng [48], who use a relative-entropy approach to prove a more constructive variant of the Min-Max theorem that leads to an asymptotically optimal dense model theorem.

By $\mathrm{Conv}(\mathcal{F})$ we denote the convex hull of $\mathcal{F}$, and by $\mathrm{Avg}_k(\mathcal{F})$ the family of all averages of at most k elements of $\mathcal{F}$. It will also be convenient to define, for any $t \in \mathbb{R}$, the threshold function $\mathrm{Th}_t : \mathbb{R} \to \{0,1\}$ by $\mathrm{Th}_t(x) = 1$ if $x \geq t$, and $\mathrm{Th}_t(x) = 0$ if $x < t$.

Let us now restate Theorem 6.1 in the following fashion.

Theorem 6.2 (Transference principle, II)**.** *Let $\epsilon, \delta > 0$. Let $\mathcal{F} = \{\phi : X \to [0,1]\}$ be a family of functions and $\nu : X \to \mathbb{R}^+$ any measure, and let $f : X \to \mathbb{R}^+$ be any function with $0 \leq f \leq \nu$ and $\mathbb{E}f = \delta$. Suppose that for any bounded function $g : X \to [0,1]$ such that $\mathbb{E}g = \delta$ we have $\|f - g\|_{\mathcal{F}} \geq \epsilon\delta$.*
Then ν is not pseudorandom in the following sense:

(1) *There exists an integer $k = O(\epsilon^{-2}\log(\epsilon^{-1}\delta^{-1}))$, $t \in \mathbb{R}$ and $\psi = \mathrm{Th}_t(\overline{\phi})$ for some $\overline{\phi} \in \mathrm{Avg}_k(\mathcal{F})$ such that*

$$\langle \psi, \nu - 1_X \rangle = \Omega(\epsilon\delta).$$

(2) *There exists $k = \mathrm{poly}(\epsilon^{-1}, \delta^{-1})$ and $\phi \in \mathcal{F}^k$ such that*

$$\langle \phi, \nu - 1_X \rangle \geq \exp(-\mathrm{poly}(\epsilon^{-1}, \delta^{-1})).$$

Part (2) immediately implies Theorem 6.1, and is itself derived from (1) by an application of the Bolzano-Weierstrass theorem, which allows us to replace the threshold function with a product of functions from $\mathcal{F}$. We refer the reader to [35] for details of this reduction, and focus on Part (1) below.

The driving force behind the proof of Theorem 6.2 (1) will be the Min-Max theorem from game theory, which asserts, roughly speaking, that in any zero-sum game between two players, if Player 2 can respond to any (potentially randomised) strategy of Player 1 and achieve a payoff of at least c, then Player 2 in fact has a *universal* (randomised) strategy that guarantees such a payoff regardless of Player 1's strategy.

Theorem 6.3 (Min-Max theorem)**.** *Consider a 2-player zero-sum game in which Player 1's set of pure strategies is $\mathcal{V} \subseteq \Delta_X$ and Player 2's is $\mathcal{W}$, and the expected pay-off to Player 2 is $\mathbb{E}f(V,W)$ for some function $f : X \times \mathcal{W} \to [0,1]$. Then*

$$\max_{W \in \mathrm{Conv}(\mathcal{W})} \min_{V \in \mathcal{V}} \mathbb{E}f(V,W) = \min_{V \in \mathrm{Conv}(\mathcal{V})} \max_{W \in \mathcal{W}} \mathbb{E}f(V,W).$$

In particular, there exist mixed strategies $V^* \in \mathrm{Conv}(\mathcal{V})$ for Player 1 and $W^* \in \mathrm{Conv}(\mathcal{W})$ for Player 2, and a value c such that

$$\min_{V \in \mathcal{V}} \mathbb{E}f(V, W^*) = \max_{W \in \mathcal{W}} \mathbb{E}f(V^*, W) = c.$$

Having stated the essential ingredient, we shall now begin a proof of Part (1) of Theorem 6.1 following [35]. First, we give some intuition behind the general argument. Let $\mathcal{H}$ be the set of all functions $g : X \to [0,1]$ such that $\mathbb{E}g = \delta$. Suppose that for all $h \in \mathcal{H}$, there

exists $\phi \in \mathcal{F}$ such that

$$|\langle f - h, \phi \rangle| > \epsilon\delta. \tag{6.1}$$

We must conclude that $\nu - 1_X$ cannot be $\epsilon\delta$-pseudorandom with respect to $\mathcal{F}$. In order to see this, suppose for the purpose of simplifying the argument that there is a pair h, ϕ as above such that $h(x) = 1$ for all $x \in \mathrm{supp}(\phi)$. Then, since $f \leq \nu$ and by the aforementioned property of h,

$$\langle \nu - 1_X, \phi \rangle = \langle \nu, \phi \rangle - \langle 1_X, \phi \rangle \geq \langle f, \phi \rangle - \langle 1_X, \phi \rangle = \langle f, \phi \rangle - \langle h, \phi \rangle = \langle f - h, \phi \rangle, \tag{6.2}$$

which in conjunction with (6.1) says that $\nu - 1_X$ is not $\epsilon\delta$-pseudorandom with respect to $\mathcal{F}$. Since we had assumed h to be a of a special form, this conclusion is not quite valid in general, but we shall see that the general approach can be made to work with a little more effort. In what follows it will be convenient to replace $\mathcal{F}$ by $\overline{\mathcal{F}} := \mathcal{F} \cup (1 - \mathcal{F})$, as this allows us to remove the absolute value in (6.1). Starting with that equation, we make use of the full strength of the Min-Max theorem to establish the existence of a *universal* distinguisher ([35], Claim 2.1).

Lemma 6.4. *There exists a function $\tilde{\phi} \in \mathrm{Conv}(\overline{\mathcal{F}})$ such that for all $h \in \mathcal{H}$, $\langle f - h, \tilde{\phi} \rangle > \epsilon\delta$.*

Proof. We consider the following 2-player zero-sum game. The first player picks $g \in \mathcal{V} := \delta^{-1}\mathcal{H} \subseteq \Delta_X$, the second picks $\phi \in \mathcal{W} := \overline{\mathcal{F}}$. The payoff for Player 1 is $-\langle \phi, \delta^{-1}f - g \rangle$, and the payoff for Player 2 is $\langle \phi, \delta^{-1}f - g \rangle$. Now by the Min-Max theorem (Theorem 6.3), the game has a value c for which Player 1 has an optimal mixed strategy $g^* \in \mathrm{Conv}(\delta^{-1}\mathcal{H})$ and Player 2 has an optimal mixed strategy $\phi^* \in \mathrm{Conv}(\overline{\mathcal{F}})$ in the sense that for all $\phi \in \overline{\mathcal{F}}$

$$\langle \phi, \delta^{-1}f - g^* \rangle \leq c, \tag{6.3}$$

and for all $g \in \delta^{-1}\mathcal{H}$,

$$\langle \phi^*, \delta^{-1}f - g \rangle \geq c \tag{6.4}$$

Since g^* is a distribution over elements of $\delta^{-1}\mathcal{H}$, it is itself in $\delta^{-1}\mathcal{H}$ and hence by hypothesis of the theorem, there exists $\phi \in \overline{\mathcal{F}}$ such that $\langle \phi, \delta^{-1}f - g^* \rangle > \epsilon$. It follows from (6.3) that $c > \epsilon$, and from (6.4) that $\langle \phi^*, \delta^{-1}f - g \rangle \geq c > \epsilon$ for all $g \in \delta^{-1}\mathcal{H}$, or $\langle \phi^*, f - g \rangle > \epsilon\delta$ for all $g \in \mathcal{H}$. □

Note that since ϕ^* was optimal, $\tilde{\phi}$ is a distribution over functions in $\overline{\mathcal{F}}$. It is therefore reasonable to expect to be able to replace this universal distinguisher $\tilde{\phi}$ by an average of a small number of functions in $\mathcal{F}$. To this end, let S be a set of the $\delta|X|$ elements of X

for which the value of $\tilde{\phi}$ is largest. Since S has density δ in X, that is $1_S \in \mathcal{H}$, we have $\langle f - 1_S, \tilde{\phi}\rangle > \epsilon\delta$. We shall show that f and 1_S can also be distinguished by a (Boolean) threshold function.

Claim 6.5. *There exists a value of $t \in [\epsilon/2, 1]$ such that*

$$\langle f, \mathrm{Th}_t(\tilde{\phi})\rangle - \langle 1_S, \mathrm{Th}_{t-\epsilon/2}(\tilde{\phi})\rangle > \epsilon\delta/2.$$

Proof of Claim 6.5: Note that by definition of the threshold function Th_t,

$$\int_0^1 \mathrm{Th}_t(\tilde{\phi})(x)dt = \tilde{\phi}(x),$$

and thus by hypothesis

$$\int_0^1 \langle f, \mathrm{Th}_t(\tilde{\phi})\rangle dt = \langle f, \tilde{\phi}\rangle > \langle 1_S, \tilde{\phi}\rangle + \epsilon\delta = \int_0^1 \langle 1_S, \mathrm{Th}_t(\tilde{\phi})\rangle dt + \epsilon\delta. \tag{6.5}$$

Suppose the statement of the claim is false, and we have that for all $t \in [\epsilon/2, 1]$,

$$\langle f, \mathrm{Th}_t(\tilde{\phi})\rangle - \langle 1_S, \mathrm{Th}_{t-\epsilon/2}(\tilde{\phi})\rangle \leq \epsilon\delta/2.$$

Then

$$\int_0^1 \langle f, \mathrm{Th}_t(\tilde{\phi})\rangle dt = \int_0^{\epsilon/2} \langle f, \mathrm{Th}_t(\tilde{\phi})\rangle dt + \int_{\epsilon/2}^1 \langle f, \mathrm{Th}_t(\tilde{\phi})\rangle dt$$

is less than or equal to

$$\int_0^{\epsilon/2} \langle f, 1_X\rangle dt + \int_{\epsilon/2}^1 (\langle 1_S, \mathrm{Th}_{t-\epsilon/2}(\tilde{\phi})\rangle + \epsilon\delta/2)dt$$

which in turn is bounded above by

$$\epsilon/2 \cdot \delta + \int_0^1 \langle 1_S, \mathrm{Th}_t(\tilde{\phi})\rangle dt + \epsilon\delta/2 \leq \int_0^1 \langle 1_S, \mathrm{Th}_t(\tilde{\phi})\rangle dt + \epsilon\delta,$$

contradicting the assumption that $\tilde{\phi}$ is capable of distinguishing f and 1_S in (6.5). □

We shall next show that $\tilde{\phi}$, and its derived Boolean threshold, are also capable of distinguishing ν and 1_X. First, let $t' := t - \epsilon/2$ and observe that the function 1_S equals 1 on $\mathrm{supp}(\mathrm{Th}_{t'}(\tilde{\phi}))$. Indeed, suppose this were not the case and we had $1_S(x) < 1$ for some $x \in \mathrm{supp}(\mathrm{Th}_{t'}(\tilde{\phi}))$. Then since the set S was defined to contain the $\delta|X|$ elements of X for which the value of $\tilde{\phi}$ is largest, we would certainly have $1_S(y) = 0$ for all $y \notin \mathrm{supp}(\mathrm{Th}_{t'}(\tilde{\phi}))$, in which case

$$\langle 1_S, \mathrm{Th}_{t'}(\tilde{\phi})\rangle = \langle 1_S, 1_X\rangle = \delta = \langle f, 1_X\rangle \geq \langle f, \mathrm{Th}_t(\tilde{\phi})\rangle,$$

contradicting Claim 6.5. Note that we are now in possession of the simplifying assumption made in the intuitive outline of the proof, preceding (6.2): we have a function which has density δ in X, namely 1_S, and a threshold function $\mathrm{Th}_{t'}(\tilde{\phi})$ on whose support 1_S equals 1, and which distinguishes f and 1_S. It follows, arguing as in (6.2), that

$$\langle \nu - 1_X, \mathrm{Th}_{t'}(\tilde{\phi})\rangle = \langle \nu, \mathrm{Th}_{t'}(\tilde{\phi})\rangle - \langle 1_X, \mathrm{Th}_{t'}(\tilde{\phi})\rangle \geq \langle f, \mathrm{Th}_{t'}(\tilde{\phi})\rangle - \langle 1_S, \mathrm{Th}_{t'}(\tilde{\phi})\rangle,$$

which is bounded below by

$$\langle f, \mathrm{Th}_{t}(\tilde{\phi})\rangle - \langle 1_S, \mathrm{Th}_{t'}(\tilde{\phi})\rangle \geq \epsilon\delta/2$$

by Claim 6.5. So again, the threshold function $\mathrm{Th}_{t'}(\tilde{\phi})$ distinguishes ν and 1_X. The additional slack in the statement will now allow us to replace the threshold function $\mathrm{Th}_{t'}(\tilde{\phi})$ with the threshold of a function that is an average of at most k functions in $\overline{\mathcal{F}}$ (rather than an element of $\mathrm{Conv}(\overline{\mathcal{F}})$).

Lemma 6.6. *Under the assumption of (6.1) and with the earlier value of t, there exists a distinguisher $\overline{\phi} \in \mathrm{Avg}_k(\mathcal{F})$ such that*

$$\langle \mathrm{Th}_{t-2\epsilon/5}(\overline{\phi}), \nu - 1_X\rangle = \Omega(\epsilon\delta).$$

Proof. We view $\tilde{\phi}$ as a distribution over functions in $\overline{\mathcal{F}}$. It follows from a Chernoff bound that $\tilde{\phi}$ will be well approximated by the average of a few functions sampled randomly from this distribution. To see this, pick k functions $\phi_1, \phi_2, \ldots, \phi_k$ randomly and independently from $\overline{\mathcal{F}}$ with probability given by the distribution $\tilde{\phi}$. Then for $k = O(\epsilon^{-2}\log(\epsilon^{-1}\delta^{-1}))$, we have that for every fixed element y,

$$\mathbb{P}_{\phi_1,\phi_2,\ldots,\phi_k}\left[\left|\tilde{\phi}(y) - \frac{\phi_1(y)+\phi_2(y)+\cdots+\phi_k(y)}{k}\right| > \epsilon/10\right] \leq \epsilon\delta/100.$$

Therefore for any probability distribution Y, we find that

$$\mathbb{E}_{\phi_1,\phi_2,\ldots,\phi_k}\left[\mathbb{P}_{y\in Y}\left[\left|\tilde{\phi}(y) - \frac{\phi_1(y)+\phi_2(y)+\cdots+\phi_k(y)}{k}\right| > \epsilon/10\right]\right] \leq \epsilon\delta/100,$$

and by Markov's inequality

$$\mathbb{P}_{\phi_1,\phi_2,\ldots,\phi_k}\left[\mathbb{P}_{y\in Y}\left[\left|\tilde{\phi}(y) - \frac{\phi_1(y)+\phi_2(y)+\cdots+\phi_k(y)}{k}\right| > \epsilon/10\right] > \epsilon\delta/10\right] \leq 1/10$$

when y is drawn from any distribution Y. It follows that there exists a choice of $\phi_1, \phi_2, \ldots, \phi_k$ such that, upon letting $\overline{\phi} := \frac{1}{k}\sum_{i=1}^k \phi_i$, we have

$$\mathbb{P}_{y\in Y}\left[|\tilde{\phi}(y) - \overline{\phi}(y)| > \epsilon/10\right] \leq \epsilon\delta/10$$

whenever y is chosen according to ν or to 1_X. In particular, this implies that

$$\langle \mathrm{Th}_{t-\epsilon/10}(\overline{\phi}), \nu\rangle \geq \langle \mathrm{Th}_t(\tilde{\phi}), \nu\rangle - \epsilon\delta/5 \geq \langle \mathrm{Th}_t(\tilde{\phi}), f\rangle - \epsilon\delta/5$$

and

$$\langle \mathrm{Th}_{t-2\epsilon/5}(\overline{\phi}), 1_X\rangle \leq \langle \mathrm{Th}_{t-\epsilon/2}(\tilde{\phi}), 1_X\rangle + \epsilon\delta/10 = \langle \mathrm{Th}_{t-\epsilon/2}(\tilde{\phi}), 1_S\rangle + \epsilon\delta/10.$$

This means that

$$\langle \mathrm{Th}_{t-2\epsilon/5}(\overline{\phi}), \nu - 1_X\rangle \geq \langle \mathrm{Th}_{t-\epsilon/10}(\overline{\phi}), \nu\rangle - \langle \mathrm{Th}_{t-2\epsilon/5}(\overline{\phi}), 1_X\rangle$$

is bounded below by

$$\langle \mathrm{Th}_t(\tilde{\phi}), f\rangle - \epsilon\delta/5 - (\langle \mathrm{Th}_{t-\epsilon/2}(\tilde{\phi}), 1_S\rangle + \epsilon\delta/10),$$

which by Claim 6.5 is $\Omega(\epsilon\delta)$. It follows that $\mathrm{Th}_{t-2\epsilon/5}(\overline{\phi})$ distinguishes ν and 1_X, as desired. □

This concludes the proof of Part (1) of Theorem 6.2. We now turn our attention to the aforementioned work of Vadhan and Zheng [48], as a consequence of which we shall be able to prove a version of Theorem 6.2 in which the parameter k can be taken to be $O(\epsilon^{-2}\log(\delta^{-1}))$. This strengthened version of the dense model theorem had previously been proved by Zhang [50], who also showed that this dependence of k on δ and ϵ is asymptotically optimal (see also [29]).

Vadhan and Zheng's main innovation is a more constructive, uniform[4] Min-Max theorem which we state and prove below. In fact, it is quite likely that an optimal dense model theorem could be obtained by applying the relative-entropy method directly, rather than to the main ingredient in the proof. However, given that the uniform version of the Min-Max theorem has numerous other applications we have preferred to include it as stated in [48].

Theorem 6.7 (Uniform Min-Max theorem)**.** *Let $m \in \mathbb{R}^+$. Consider a 2-player zero-sum game in which Player 1's set of pure strategies is $\mathcal{V} \subseteq \Delta_X$ and Player 2's is $\mathcal{W}$, and the expected pay-off to Player 2 is $\mathbb{E}f(V,W)$ for some function $f : X \times \mathcal{W} \to [-m,m]$.*
Then for all $0 < \epsilon < 1$ there exists an integer S and an algorithm which produces a sequence of strategies $V^{(1)}, V^{(2)}, \dots, V^{(S)} \in \mathrm{Conv}(\mathcal{V})$ and $W^{(1)}, W^{(2)}, \dots, W^{(S)} \in \mathcal{W}$ and outputs a mixed strategy $W^* \in \mathrm{Conv}(\mathcal{W})$ such that

$$\mathbb{E}f(V,W^*) \geq \mathrm{avg}_{1\leq i\leq S}\mathbb{E}f(V^{(i)},W^{(i)}) - O(m\epsilon).$$

[4]The terminology "uniform" takes its meaning from applications of the Min-Max theorem in cryptography, where one seeks to construct an adversary "uniformly" (see [52]).

for any Player 1 strategy $V \in \mathcal{V}$, where S satisfies $D_{KL}(V||V^{(1)}) \leq S\epsilon^2$ for all $V \in \mathcal{V}$.

We first show that Theorem 6.7 implies the original Min-Max Theorem.

PROOF OF THEOREM 6.3 USING THEOREM 6.7: Since $W^{(i)}$ in Theorem 6.7 will be defined to be Player 2's best response to the mixed strategy $V^{(i)}$, we have

$$\mathbb{E}f(V^{(i)}, W^{(i)}) = \max_{W \in \mathcal{W}} \mathbb{E}f(V^{(i)}, W).$$

By Theorem 6.7, there exists a strategy $W^* \in \mathrm{Conv}(\mathcal{W})$ such that

$$\min_{V \in \mathcal{V}} \mathbb{E}f(V, W^*) \geq \mathrm{avg}_{1 \leq i \leq S} \mathbb{E}f(V^{(i)}, W^{(i)}) - O(m\epsilon) = \mathrm{avg}_{1 \leq i \leq S} \max_{W \in \mathcal{W}} \mathbb{E}f(V^{(i)}, W) - O(m\epsilon).$$

But $\max_{W \in \mathcal{W}} \mathbb{E}f(V^{(i)}, W) \geq \min_{V \in \mathrm{Conv}(\mathcal{V})} \max_{W \in \mathcal{W}} \mathbb{E}f(V, W)$ for all i, so

$$\max_{W \in \mathrm{Conv}(\mathcal{W})} \min_{V \in \mathcal{V}} \mathbb{E}f(V, W) \geq \min_{V \in \mathcal{V}} \mathbb{E}f(V, W^*) \geq \min_{V \in \mathrm{Conv}(\mathcal{V})} \max_{W \in \mathcal{W}} \mathbb{E}f(V, W) - O(m\epsilon).$$

Letting $\epsilon \to 0$ yields the statement of Theorem 6.3, the reverse inequality being trivial. □

The proof of Theorem 6.7 proceeds by an iterative procedure which incrementally decreases the relative entropy of $V \in \mathcal{V}$ using multiplicative weight updates analogous to the proof of Theorem 3.2 in Section 2. Vadhan and Zheng credit the idea to Barak, Hardt and Kale, who used multiplicative weight updates coupled with approximate KL-projections to give a simple, more efficient and uniform proof of the hard-core lemma [1].

Definition 6.8 (KL projection)**.** *Let Z be a distribution on X and let $\mathcal{V}$ be any non-empty closed convex set of distributions on X. We say $\widehat{Y} \in \mathcal{V}$ is a* KL projection of Z onto $\mathcal{V}$ *if*

$$\widehat{Y} = \arg\min_{Y \in \mathcal{V}} D_{KL}(Y||Z).$$

KL projections satisfy the following Pythagorean property (see [6], Chapter 11).

Proposition 6.9 (Pythagorean property)**.** *Let $\mathcal{V}$ be any non-empty closed convex set of distributions on X, and let $\widehat{Y} \in \mathcal{V}$ be a KL projection of Z onto $\mathcal{V}$. Then for all $y \in \mathcal{V}$,*

$$D_{KL}(Y||\widehat{Y}) + D_{KL}(\widehat{Y}||Z) \leq D_{KL}(Y||Z).$$

It is easy to see that the Pythagorean property implies that the KL projection is unique. However, finding the exact KL projection is often not feasible, so we make do with the following approximate notion.

Definition 6.10 (Approximate KL projection)**.** *Let $0 < \eta < 1$, let Z be a distribution on X and let $\mathcal{V}$ be any non-empty closed convex set of distributions on X. We say $\widetilde{Y}$ is an*

η-approximate KL projection of Z onto $\mathcal{V}$ *if* $\widetilde{Y} \in \mathcal{V}$ *and, for all* $Y \in \mathcal{V}$,

$$D_{KL}(Y||\widetilde{Y}) \leq D_{KL}(Y||Z) + \eta.$$

Note that if

$$D_{KL}(Y||\widetilde{Y}) \leq D_{KL}(Y||\widehat{Y}) + \eta$$

for all $Y \in \mathcal{V}$, where $\widehat{Y}$ is the (exact) KL projection of Z onto $\mathcal{V}$, then $\widetilde{Y}$ is an η-approximate KL projection of Z onto $\mathcal{V}$. We are now ready to state the iterative procedure leading to a proof of Theorem 6.7.

Procedure 6.11. *Apply the following iterative procedure.*

Start the algorithm with an arbitrary $V^{(1)} \in \mathrm{Conv}(\mathcal{V})$. *For* $i = 1, 2, \ldots, S$,

- *let* $W^{(i)} \in \mathcal{W}$ *be a best Player-2 response to* $V^{(i)}$;
- *let* $V^{(i)'}$ *be such that*

$$\mathbb{P}[V^{(i)'} = x] \propto \exp(-\epsilon f(x, W^{(i)})/2m) \cdot \mathbb{P}[V^{(i)} = x];$$

- *let* $V^{(i+1)}$ *be an arbitrary* ϵ^2*-approximate KL projection of* $V^{(i)'}$ *onto* $\mathrm{Conv}(\mathcal{V})$.

Output the mixed strategy

$$W^* := \frac{1}{S}\sum_{i=1}^{S} W^{(i)}.$$

The following lemma states that such multiplicative weight updates decrease KL divergence.

Lemma 6.12. *Let* A, B *be distributions over* X *and* $h : X \to [0, 1]$ *be any function. Define a distribution* A' *such that*

$$\mathbb{P}[A' = x] \propto \exp(\epsilon h(x))\mathbb{P}[A = x]$$

for $0 \leq \epsilon \leq 1$. *Then*

$$D_{KL}(B||A') \leq D_{KL}(B||A) - (\log e)\epsilon(\mathbb{E}[h(B)] - \mathbb{E}[h(A)] - \epsilon).$$

Proof. By definition,

$$D_{KL}(B||A) - D_{KL}(B||A') = \sum_x \mathbb{P}[B = x] \log \frac{\mathbb{P}[A' = x]}{\mathbb{P}[A = x]} = \sum_x \mathbb{P}[B = x] \log \frac{e^{\epsilon h(x)}}{\sum_y e^{\epsilon h(y)}\mathbb{P}[A = y]},$$

which equals

$$(\log e)\left(\epsilon \mathbb{E}h(B) - \ln \sum_y e^{\epsilon h(y)}\mathbb{P}[A = y]\right).$$

Bounding the exponential above and below by $e^z \leq 1+z+z^2$ and $1+z \leq e^z$, respectively, we find that

$$\begin{aligned} D_{KL}(B||A) - D_{KL}(B||A') &\geq (\log e)\left(\epsilon\mathbb{E}[h(B)] - \ln(1+\epsilon\mathbb{E}[h(A)] + \epsilon^2)\right) \\ &\geq (\log e)\epsilon(\mathbb{E}[h(B)] - \mathbb{E}[h(A)] - \epsilon) \end{aligned}$$

as desired. □

Proof of Theorem 6.7: Let S be the least integer such that for any $V \in \mathcal{V}$, $D_{KL}(V||V^{(1)}) \leq S\epsilon^2$. By Lemma 6.12 with $A = V^{(i)}$, $A' = V^{(i)'}$, $B = V$ and $h(x) = -f(x, W^{(i)})/2m$, we have

$$D_{KL}(V||V^{(i)}) - D_{KL}(V||V^{(i)'}) \geq (\log e)\epsilon\left(\frac{\mathbb{E}f(V^{(i)}, W^{(i)}) - \mathbb{E}f(V, W^{(i)})}{2m} - \epsilon\right).$$

Now since $V^{(i+1)}$ is an ϵ^2-approximate KL projection of $V^{(i)'}$ onto $\mathrm{Conv}(\mathcal{V})$, we have $D_{KL}(V||V^{(i+1)}) \leq D_{KL}(V||V^{(i)'}) + \epsilon^2$ and thus

$$D_{KL}(V||V^{(i)}) - D_{KL}(V||V^{(i+1)}) \geq (\log e)\epsilon\left(\frac{\mathbb{E}f(V^{(i)}, W^{(i)}) - \mathbb{E}f(V, W^{(i)})}{2m} - \epsilon\right) - \epsilon^2.$$

Summing the telescoping series from 1 to S, we have

$$D_{KL}(V||V^{(1)}) - D_{KL}(V||V^{(S+1)}) \geq (\log e)\epsilon\sum_{i=1}^{S}\left(\frac{\mathbb{E}f(V^{(i)}, W^{(i)}) - \mathbb{E}f(V, W^{(i)})}{2m} - \epsilon\right) - S\epsilon^2,$$

with the latter expression being equal to

$$(\log e)S\epsilon\left(\frac{\mathrm{avg}_{1\leq i\leq S}\mathbb{E}f(V^{(i)}, W^{(i)}) - \mathbb{E}f(V, W^*)}{2m} - \epsilon\right) - S\epsilon^2.$$

Since $D_{KL}(V||V^{(S+1)}) \geq 0$, it follows that

$$\frac{\mathrm{avg}_{1\leq i\leq S}\mathbb{E}f(V^{(i)}, W^{(i)}) - \mathbb{E}f(V, W^*)}{2m} \leq \frac{D_{KL}(V||V^{(1)}) + S\epsilon^2}{(\log e)S\epsilon} + \epsilon = O(\epsilon),$$

which proves the theorem. □

Vadhan and Zheng [48] use the Uniform Min-Max theorem to strengthen and reprove a number of known results in complexity theory. For example, they give a new proof of Impagliazzo's hard core theorem [25] with optimal hard core density and optimal complexity blow-up. They also deduce an optimal weak regularity lemma for graphs of density $o(1)$

([48], Section 6.2). Here we shall focus on showing how Theorem 6.7 gives a quick proof of the following version of the dense model theorem due to Zhang [50].

Theorem 6.13 (Optimal dense model theorem)**.** *Let $\epsilon, \delta > 0$. Let $\mathcal{F} = \{\phi : X \to [0,1]\}$ be any family of functions and $\nu : X \to \mathbb{R}^+$ any measure, and let $f : X \to \mathbb{R}^+$ be any function with $0 \leq f \leq \nu$ and $\mathbb{E}f = \delta$. Suppose that for any bounded function $g : X \to [0,1]$ such that $\mathbb{E}g = \delta$, we have $\|f - g\|_{\mathcal{F}} \geq \epsilon\delta$.*
Then ν is not pseudorandom in the sense that there exists an integer $k = O(\epsilon^{-2}\log(\delta^{-1}))$, $t \in \mathbb{R}$ and $\psi = \mathrm{Th}_t(\overline{\phi})$ for some $\overline{\phi} \in \mathrm{Avg}_k(\mathcal{F})$ such that

$$\langle \psi, \nu - 1_X \rangle = \Omega(\epsilon\delta).$$

Proof. Note that in the proof of Theorem 6.2, we only actually used the inequality

$$c =: \min_{V \in \mathcal{V}} \mathbb{E}f(V, W^*) \geq \max_{W \in \mathcal{W}} \mathbb{E}f(V^*, W),$$

which we now replace with

$$c = \min_{V \in \mathcal{V}} \mathbb{E}f(V, W^*) \geq \mathrm{avg}_{1 \leq i \leq S} \mathbb{E}f(V^{(i)}, W^{(i)}) - O(\epsilon)$$

as a result of applying Theorem 6.7 with $m = 1$, $\mathcal{V} = \delta^{-1}\mathcal{H} \subseteq \Delta_X$ and $\mathcal{W} = \mathcal{F}$. Note that since $\mathcal{V}$ contains the uniform distribution on X, and we may start the algorithm with an arbitrary $V^{(1)} \in \mathcal{V}$, we may choose S to be the least integer such that for all $V \in \mathcal{V}$, $D_{KL}(V||V^{(1)}) = D_{KL}(V||U_X) = \mathrm{Ent}(V) \leq S\epsilon^2$, which means S can be taken to be $O(\epsilon^{-2}\log(\delta^{-1}))$. We proceed almost exactly as in the proof of Lemma 6.4 to obtain our universal distinguisher. Indeed, note that by hypothesis, for every $V \in \mathcal{V}$ there exists $W \in \mathcal{W}$ such that $\mathbb{E}f(V, W) \geq \epsilon$. This is true in particular for every $V^{(i)}$, implying that

$$\mathrm{avg}_{1 \leq i \leq S} \mathbb{E}f(V^{(i)}, W^{(i)}) = \mathrm{avg}_{1 \leq i \leq S} \max_{W \in \mathcal{W}} \mathbb{E}f(V^{(i)}, W) \geq \epsilon,$$

so that as before, $c \geq \epsilon$. We conclude that there exists a universal distinguisher ϕ^* such that for all $g \in \mathcal{H}$,

$$\langle \phi^*, f - g \rangle \geq \epsilon\delta,$$

but by definition of Procedure 6.11, the distinguisher ϕ^* is already an average of S elements of $\overline{\mathcal{F}}$. □

It would be interesting to determine whether Theorem 6.13 has any direct applications in additive combinatorics. Regardless of the answer to this particular question, there is little doubt that the general approach using relative entropy outlined in this article will find numerous further uses.

References

[1] Boaz Barak, Moritz Hardt, and Satyen Kale. The uniform hardcore lemma via approximate Bregman projections. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1193–1200. SIAM, Philadelphia, PA, 2009.

[2] Thomas Bloom. A quantitative improvement for Roth's theorem on arithmetic progressions. *J. Lond. Math. Soc. (2)*, 93(3):643–663, 2016.

[3] Mei-Chu Chang. A polynomial bound in Freiman's theorem. *Duke Math. J.*, 113(3):399–419, 2002.

[4] David Conlon, Jacob Fox, and Yufei Zhao. The Green-Tao theorem: an exposition. *EMS Surv. Math. Sci.*, 1(2):249–282, 2014.

[5] David Conlon, Jacob Fox, and Yufei Zhao. A relative Szemerédi theorem. *Geom. Funct. Anal.*, 25(3):733–762, 2015.

[6] Thomas Cover and Joy Thomas. *Elements of information theory*. Wiley Series in Telecommunications. John Wiley & Sons, Inc., New York, New York, USA, 1991.

[7] Jacob Fox. A new proof of the graph removal lemma. *Ann. of Math. (2)*, 174(1):561–579, 2011.

[8] Ehud Friedgut. An information-theoretic proof of a hypercontractive inequality. *arXiv*, 1504.01506, April 2015.

[9] Alan Frieze and Ravi Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999.

[10] David Galvin. Three tutorial lectures on entropy and counting. *arXiv*, 1406.7872, June 2014.

[11] Timothy Gowers. A new proof of Szemerédi's theorem for arithmetic progressions of length four. *Geom. Funct. Anal.*, 8(3):529–551, 1998.

[12] Timothy Gowers. A new proof of Szemerédi's theorem. *Geom. Funct. Anal.*, 11(3):465–588, 2001.

[13] Timothy Gowers. Decompositions, approximate structure, transference, and the Hahn-Banach theorem. *Bull. Lond. Math. Soc.*, 42(4):573–606, 2010.

[14] Timothy Gowers. Entropy and Sidorenko's conjecture — after Szegedy. *Personal blog*, https://gowers.wordpress.com/2015/11/18/entropy-and-sidorenkos-conjecture-after-szegedy/, November 2015.

[15] Timothy Gowers and Julia Wolf. Linear forms and quadratic uniformity for functions on $\mathbb{F}_p^n$. *Mathematika*, 57(2):215–237, 2012.

[16] Ben Green. Some constructions in the inverse spectral theory of cyclic groups. *Combin. Probab. Comput.*, 12(2):127–138, 2003.

[17] Ben Green. Spectral structure of sets of integers. In *Fourier analysis and convexity*, pages 83–96. Birkhäuser Boston, Boston, MA, 2004.

[18] Ben Green. Finite field models in additive combinatorics. In Bridget S Webb, editor, *Surveys in combinatorics 2005*, pages 1–27. Cambridge Univ. Press, Cambridge, Cambridge, 2005.

[19] Ben Green. Roth's theorem in the primes. *Ann. of Math. (2)*, 161(3):1609–1636, 2005.

[20] Ben Green. Montréal notes on quadratic Fourier analysis. In *Additive combinatorics*, pages 69–102. Amer. Math. Soc., Providence, RI, 2007.

[21] Ben Green and Tom Sanders. Boolean functions with small spectral norm. *Geom. Funct. Anal.*, 18(1):144–162, 2008.

[22] Ben Green and Terence Tao. An inverse theorem for the Gowers $U^3(G)$ norm. *Proc. Edinb. Math. Soc. (2)*, 51(1):73–153, 2008.

[23] Ben Green and Terence Tao. The primes contain arbitrarily long arithmetic progressions. *Ann. of Math. (2)*, 167(2):481–547, 2008.

[24] Ben Green and Terence Tao. Linear equations in primes. *Ann. of Math. (2)*, 171(3):1753–1850, 2010.

[25] Russell Impagliazzo. Hard-core distributions for somewhat hard problems. In *FOCS '07. 48th Annual IEEE Symposium on Foundations of Computer Science, 2007.*, pages 538–545. IEEE, 1995.

[26] Russell Impagliazzo, Cristopher Moore, and Alexander Russell. An entropic proof of Chang's inequality. *SIAM J. Discrete Math.*, 28(1):173–176, 2014.

[27] Solomon Kullback and Richard Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22(1):79–86, 1951.

[28] James Lee. Covering the large spectrum and generalized Riesz products. *arXiv*, 1508.07109, August 2015.

[29] Chi-Jen Lu, Shi-Chun Tsai, and Hsin-Lung Wu. Complexity of hard-core set proofs. *Comput. Complexity*, 20(1):145–171, 2011.

[30] Guy Moshkovitz and Asaf Shapira. A sparse regular approximation lemma. *arXiv*, 1610.02676, October 2016.

[31] Ryan O'Donnell. Lecture 16: The hypercontractivity theorem. *Personal website*, https://www.cs.cmu.edu/~odonnell/boolean-analysis/lecture16.pdf.

[32] Sean Prendiville. Four variants of the Fourier-analytic transference principle. *arXiv*, 1509.09200, September 2015.

[33] Jaikumar Radhakrishnan. An entropy proof of Bregman's theorem. *J. Combin. Theory Ser. A*, 77(1):161–164, 1997.

[34] Omer Reingold, Luca Trevisan, Madhur Tulsiani, and Salil Vadhan. Dense subsets of pseudorandom sets. *Electronic Colloquium on Computational Complexity*, 45:1–33, 2008.

[35] Omer Reingold, Luca Trevisan, Madhur Tulsiani, and Salil Vadhan. New proofs of the Green-Tao-Ziegler dense model theorem: an exposition. *arXiv*, 0806.0381, June 2008.

[36] Sheldon Ross. *A first course in probability.* Macmillan Publishing Co., Inc., New York; Collier Macmillan Publishers, London, 1976.

[37] Alex Samorodnitsky. Low-degree tests at large distances. In *STOC'07—Proceedings of the 39th Annual ACM Symposium on Theory of Computing*, pages 506–515. ACM, New York, New York, New York, USA, 2007.

[38] Tom Sanders. On Roth's theorem on progressions. *Ann. of Math. (2)*, 174(1):619–636, 2011.

[39] Tom Sanders. On the Bogolyubov–Ruzsa lemma. *Anal. PDE*, 5(3):627–655, 2012.

[40] Ilya Shkredov. On sets of large exponential sums. *Dokl. Math.*, 74(3):860–864, 2006.

[41] Balazs Szegedy. An information theoretic approach to Sidorenko's conjecture. *arXiv*, 1406.6738, June 2014.

[42] Terence Tao. Moser's entropy compression argument. *Personal blog*, https://terrytao.wordpress.com/2009/08/05/mosers-entropy-compression-argument/, August 2009.

[43] Terence Tao. *Higher order Fourier analysis*, volume 142 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.

[44] Terence Tao. Entropy and rare events. *Personal blog*, https://terrytao.wordpress.com/2015/09/20/entropy-and-rare-events/, September 2015.

[45] Terence Tao. The logarithmically averaged Chowla and Elliott conjectures for two-point correlations. *Forum Math. Pi*, 4:e8–36, 2016.

[46] Terence Tao and Tamar Ziegler. The primes contain arbitrarily long polynomial progressions. *Acta Math.*, 201(2):213–305, 2008.

[47] Madhur Tulsiani and Julia Wolf. Quadratic Goldreich-Levin theorems. *SIAM J. Comput.*, 43(2):730–766, 2014.

[48] Salil Vadhan and Jia Zheng. A uniform min-max theorem with applications in cryptography. In *Advances in Cryptology – CRYPTO 2013*, pages 93–110. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[49] Julia Wolf. Finite field models in arithmetic combinatorics – ten years on. *Finite Fields Appl.*, 32:233–274, 2015.

[50] Jiapeng Zhang. On the query complexity for Showing Dense Model. *Electronic Colloquium on Computational Complexity*, 38, 2011.

[51] Yufei Zhao. An arithmetic transference proof of a relative Szemerédi theorem. *arXiv*, 1307.4959, July 2013.

[52] Jia Zheng. *A uniform min-max theorem and characterizations of computational randomness*. PhD thesis, 2014.

School of Mathematics, University of Bristol, Bristol BS8 1TW
E-mail address: julia.wolf@bristol.ac.uk