Low degree almost Boolean functions are sparse juntas

Irit Dinur* Yuval Filmus† Prahladh Harsha‡

February 19, 2021

Abstract

We analyze the structure of Boolean functions on $\{0,1\}^n$ which are close to having degree d with respect to the p-biased measure, for $p \leq 1/2$. We show that such functions are close to *sparse juntas*, a new class of functions arising from our main theorem. This implies that Boolean functions which are close to having degree d are almost constant, as long as the distance from degree d is not much smaller than p. Our results extend to the *slice*, which is the subset of $\{0,1\}^n$ consisting of all vectors whose Hamming weight is pn.

When the function in question is monotone, the approximating sparse junta is a width-d monotone DNF whose coefficients are "well-spread". In the general case, instead of a disjunction of clauses we have a linear combination of clauses, whose weights belong to a finite set depending only on d.

Our main theorem is a common generalization of two known results: the Kindler–Safra theorem, which states that a Boolean function close to degree d with respect to the uniform measure is close to a junta; and the p-biased Friedgut–Kalai–Naor theorem, which is the special case $d = 1$ of our main theorem. Our methods naturally lead to a new proof of the Kindler–Safra theorem, which proceeds by induction on d.

1 Introduction

Friedgut's junta theorem [Fri98] and the Kindler–Safra theorem [Kin03] are two classical structural results in Boolean function analysis, which state that "simple" Boolean functions are close to juntas (functions depending on a constant number of coordinates), for two different measures of simplicity which we describe below. Both results hold under the uniform measure over the Boolean cube $\{0,1\}^n$.

Hatami [Hat12] generalized Friedgut's junta theorem to arbitrary product spaces, uncovering in the process a new class of functions called *pseudojuntas* which serve as the class of approximating functions for his result.

In this paper, we continue the structural study of Boolean functions on more general spaces, proving an analog of the Kindler–Safra theorem for the p-biased measure (described below), uncovering in the process a new class of functions, *sparse juntas*, which serve as the class of approximating functions for our result.

Background

Boolean function analysis [O'D14] is a topic of study at the intersection of functional analysis, combinatorics, and theoretical computer science. It studies properties of Boolean-valued or almost Boolean-valued

*Weizmann Institute of Science, ISRAEL. email: irit.dinur@weizmann.ac.il.
†Technion — Israel Institute of Technology, ISRAEL. email: yuvalfi@cs.technion.ac.il
‡Tata Institute of Fundamental Research, INDIA. email: prahladh@tifr.res.in

functions on the Boolean cube $\{0,1\}^n$ from the perspective of discrete Fourier analysis. Two classical results in the area are the two mentioned above, Friedgut's junta theorem and the Kindler–Safra theorem.

Friedgut's junta theorem [Fri98] concerns functions with low average sensitivity. The *sensitivity* of a Boolean function $f\colon \{0,1\}^n \to \{0,1\}$ at a point $x \in \{0,1\}^n$ is the number of coordinates i such that flipping the ith coordinate flips the value of the function. The *average sensitivity* of f is the expected sensitivity of a uniformly random point.

Friedgut's junta theorem states that if a Boolean function f has average sensitivity I then for every $\epsilon > 0$ there exists a Boolean function g, depending on $2^{O(I/\epsilon)}$ points, such that $\Pr[f \neq g] \leq \epsilon$ (with respect to the uniform distribution).

The Kindler–Safra theorem (and a related result of Bourgain [Bou02]) concerns functions with low degree. The *degree* of a function $f\colon \{0,1\}^n \to \{0,1\}$ can be defined in several ways: it is the degree of the unique multilinear polynomial expansion of f; and it is the minimal d such that f can be written as a linear combination of d-juntas (functions depending on d coordinates). The orthogonal projection of f to the space of (not necessarily Boolean) degree d functions is denoted by $f^{\leq d}$.

The Kindler–Safra theorem states that if a Boolean function f is close to degree d, in the sense $\mathbb{E}[(f - f^{\leq d})^2] = \epsilon$, then there exists a Boolean function g, depending on $2^{O(d)}$ coordinates, such that $\Pr[f \neq g] = O(\epsilon)$.

Biased measures While the uniform measure is the measure which is most commonly studied, in some cases other measures are more natural. A case in point is the study of the Erdős–Rényi $G(n, p)$ random graph model, in which each edge is present (independently) with probability p. The corresponding measure on the Boolean cube $\{0,1\}^{\binom{n}{2}}$ is the biased measure μ_p, given by the formula

$$\mu_p(x) = p^{\sum_i x_i}(1-p)^{\sum_i (1-x_i)}.$$

This setting is known as the *p-biased setting*.

Average sensitivity is related to the study of sharpness of thresholds of monotone graph properties via the Russo–Margulis formula, $(d/dp)\, \mathbb{E}_{\mu_p}[f] = p(1-p)I_p[f]$, where $I_p[f]$ is the μ_p-analog of average sensitivity; see for example [FK96]. A graph property has a coarse threshold if $p(1-p)I_p[f] = O(1)$ near the critical probability.

Friedgut's junta theorem fails in this setting. For example, if f is the indicator of the graph containing a triangle, then $p(1-p)I_p[f] = O(1)$, yet f cannot be approximated by a junta. Instead, Friedgut's sharp threshold theorem [Fri99] and Bourgain's related result [Bou99] describe features of f implied by the condition $p(1-p)I_p[f] = O(1)$ when f is a monotone graph property; see also [BK97]. Hatami [Hat12] proved a general structure theorem for arbitrary functions satisfying $p(1-p)I_p[f] = O(1)$, showing that they are close to *pseudojuntas*, a class of functions whose exact definition we skip.

Biased Friedgut–Kalai–Naor theorem Our goal in this paper is to generalize the Kindler–Safra theorem to the p-biased setting. A forerunner to our paper is the generalization of the Friedgut–Kalai–Naor theorem [FKN02] to the p-biased setting, due to Filmus [Fil16a].

The Friedgut–Kalai–Naor theorem [FKN02] is the case $d = 1$ of the Kindler–Safra theorem. It states that if a Boolean function f satisfies $\mathbb{E}[(f - f^{\leq 1})^2] = \epsilon$ then there is a Boolean function g *depending on a single coordinate* ("dictator") such that $\Pr[f \neq g] = O(\epsilon)$.

This theorem fails in the p-biased setting when p is small, as witnessed by the function

$$f(y_1, \ldots, y_n) = \max(y_1, \ldots, y_m), \text{ where } m = O\left(\frac{\sqrt{\epsilon}}{p}\right).$$

This function is close to the degree 1 function $h = y_1 + \cdots + y_m$. Since the distribution of h is roughly Poisson with expectation $O(\sqrt{\epsilon})$, the probability that $h \notin \{0, 1\}$ is $O(\epsilon)$, and this implies that $\mathbb{E}[(f - f^{\leq 1})]^2 \leq \mathbb{E}[(f - h)^2] = O(\epsilon)$;[1] yet f is not close to a junta, let alone a dictator!

Filmus [Fil16a] showed that in the p-biased setting for $p \leq 1/2$, if $\mathbb{E}[(f - f^{\leq 1})^2] = \epsilon$ then either f or $1 - f$ must be $O(\epsilon)$-close to the maximum of at most $\max\big(O(\sqrt{\epsilon}/p), 1\big)$ coordinates, showing that the above example is essentially the only possible one; taking a maximum with 1 is necessary to accommodate the example $f = y_1$, in which $\epsilon = 0$.

The question motivating our work is:

What kind of structure do almost degree d Boolean functions have in the p-biased setting?

The correct structure isn't clear even for $d = 2$. Here are some examples of degree 2 functions which are almost Boolean; rounding them to Boolean, we obtain Boolean functions satisfying $\mathbb{E}[(f - f^{\leq 2})^2] = \Theta(\epsilon)$:

1. Disjoint pairs: $\sum_{i=1}^{m} x_i y_i$ for $m = O(\sqrt{\epsilon}/p^2)$.

2. Non-disjoint pairs: $\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} x_i y_{i,j}$ for $m_1 m_2 = O(\sqrt{\epsilon}/p^2)$.

3. Intertwined XOR: $\sum_{i=1}^{m} y_i - 2\sum_{1 \leq i < j \leq m} y_i y_j$ for $m = O(\sqrt[3]{\epsilon}/p)$.

4. Intertwined OR: $\sum_{i=1}^{m} y_i - \sum_{1 \leq i < j \leq m} y_i y_j$ for $m = O(\sqrt[4]{\epsilon}/p)$.

Our results

Main result Our main result states that in the p-biased setting for small p, and for constant d, if a Boolean function $f\colon \{0,1\}^n \to \{0,1\}$ satisfies $\mathbb{E}[(f - f^{\leq d})^2] = \epsilon$ then $\Pr[f \neq g] = O(\epsilon)$ for some degree d *sparse junta* g, which is a function satisfying the following properties:

1. g is almost Boolean, in the sense that $\Pr[g \notin \{0,1\}] = O(\epsilon)$.

2. If we expand g as a multilinear polynomial then all coefficients belong to some finite set depending on d (when $d = 1$, this set is $\{0, \pm 1\}$).

3. On a random input, with probability $1 - O(\epsilon)$ the number of "active" monomials in g is $O(1)$; a monomial is *active* if it evaluates to 1 and appears with non-zero coefficient in the expansion of g.

Sparse juntas satisfy many more properties, listed in the statement of our main theorem, Theorem 3.1. The most important of them describe more properties of the multilinear expansion of g, formulated in terms of the *support* of g, which is the set of monomials with non-zero coefficients in the multilinear expansion of g:

1. The number of degree e monomials in the support of g is $O(\epsilon^{C_d}/p^e + 1)$, where C_d is a constant depending on d (for example, $C_1 = 1/2$).

[1] By definition, $f^{\leq 1}$ is the degree 1 function minimizing $\mathbb{E}[(f - f^{\leq 1})^2]$.

2. The number of degree e monomials in the support of g which are multiples of $y_{i_1} \cdots y_{i_t}$ is $O(1/p^{e-t})$.

Another property which follows from our main theorem is that f is $O(\epsilon^{C_d} + p)$-close to a *constant* Boolean function. We are not aware of a proof of this property that doesn't go via our main theorem.

Y-expansion One feature of sparse juntas is the quantization of the coefficients of their expansion as multilinear polynomials, an expansion which we call the *y-expansion* (it has various other names in the literature). This expansion stands in distinction to the *p-biased Fourier expansion*, in which the function is expanded as a multilinear polynomial in the variables

$$x_i := \frac{y_i - p}{\sqrt{p(1-p)}}.$$

The p-biased Fourier expansion has the advantage that the monomials $\prod_{i \in S} x_i$ form an orthonormal basis with respect to μ_p, a feature not satisfied by the monomials $\prod_{i \in S} y_i$. However, the quantization of the coefficients is only apparent in the y-expansion. As a simple example, even the function $f = y_1$ has a p-biased Fourier expansion in which the coefficients depend on p:

$$f = p + \sqrt{p(1-p)} x_i.$$

Monotone version When f is a monotone function, we can approximate f by a function g with a simpler structure: a monotone DNF of width d. A monotone DNF is a disjunction of clauses, each of which is a conjunction of variables, for example $y_1 \vee (y_2 \wedge y_3)$. The width of a DNF is the maximum number of variables in each clause (in the example, 2). The function g is also a sparse junta:

1. On a random input, with probability $1 - O(\epsilon)$ the number of clauses evaluating to true is $O(1)$.

2. The clauses comprising g satisfy the two properties stated above for the support of g.

3. Either $g \equiv 1$ or $\Pr[g \neq 0] = O(\epsilon^{C_d} + p)$.

Slice versions Both of our main results generalize to the *slice* $\binom{[n]}{k}$, which is the part of the Boolean cube $\{0,1\}^n$ consisting of all vectors of Hamming weight k. This setting appears most conspicuously in the $G(n, m)$ random graph model.

Our results hold with two necessary changes: we require k to be large enough (as a function of d), and k/n plays the role of p.

On the proof The main idea of the proof is to reduce to the Kindler–Safra theorem in the unbiased setting (that is, constant p). This is done by using the following two-step process to sample a point according to μ_p:

1. Sample a point according to μ_{2p}.

2. Resample each 1-coordinate according to $\mu_{1/2}$.

Since $2p \cdot (1/2) = p$, this results in a point distributed according to μ_p.

Here is a different way of looking at the same process. First we sample a set $S \sim \mu_{2p}$, meaning that we put each coordinate into S with probability $2p$. We then sample a point in $\{0,1\}^S$ according to $\mu_{1/2}$, and interpret it as a point in $\{0,1\}^n$ by filling the rest of the coordinates with zeroes.

For each S, we look at the restriction $f|_S$ of f to $\{0,1\}^S$, and apply the unbiased Kindler–Safra theorem to obtain a Boolean degree d function g_S close to $f|_S$. The challenge is to paste these functions g_S into a single function g.

So far we have followed the same script as in the proof of the p-biased Friedgut–Kalai–Naor theorem [Fil16a]. The main novelty of our proof is the pasting step, which uses a "higher-dimensional agreement theorem". The agreement theorem states that if the functions g_S locally agree with each other then they can be pasted to a global function g which agrees with the local functions g_S for most S.

The conference version of the paper [DFH19] used the agreement theorem proved expressly for this purpose in the companion work [DFH17]. In this version we give a self-contained proof of the special case of the agreement theorem in which the functions g_S are juntas. The proof of this special case is much easier than the proof of the general agreement theorem appearing in [DFH17], and it suffices for our purposes.

New proof of Kindler–Safra Our work has inspired us to give a new proof of the Kindler–Safra theorem, by induction on the degree. In order for the inductive argument to go through, we need to generalize the statement of the Kindler–Safra theorem: instead of considering just Boolean-valued functions, we have to consider A-valued functions for arbitrary finite A.

Our main theorem holds for A-valued functions as well, and this leads to the following succinct formulation:

If a degree d function is approximately A-valued,
then the coefficients of its y-expansion are approximately B-valued,

where B is a finite set depending on d, A.

Organization of this paper After a short preliminary section, Section 2, we prove our main theorem and its slice version in Section 3. We prove the monotone version of our main theorem and its slice version in Section 4. The proof of our main theorem requires an A-valued version of the unbiased Kindler–Safra theorem, which we prove by reduction to the Boolean case in Section 5. An alternative proof from scratch (alluded to above) is presented in Section 6. Finally, we prove the necessary agreement theorem in Section 7.

Acknowledgements We thank Dor Minzer for suggesting a monotone version of our main theorem, and for suggesting the proof of the agreement theorem, Theorem 3.9. We thank Rajat Mittal for bringing [GR97] to our attention.

2 Preliminaries

Big O We will use the notation $O_d(\cdot)$ to stress that the hidden constant depends on d.

Indicator We denote the indicator of an event $\mathcal{E}$ by $[\![\mathcal{E}]\!]$.

L2 triangle inequality We will often make use of the inequality $(a+b)^2 \leq 2(a^2+b^2)$.

Junta A function $f\colon \{0,1\}^n \to \mathbb{R}$ is a *junta* if it depends on $O(1)$ coordinates. (This somewhat informal term makes sense when the bound on the number of coordinates is clear from context.)

The measure μ_p Most of this paper concerns functions on the Boolean cube $\{0,1\}^n$, considered with respect to the biased measure μ_p. The measure μ_p is a product measure: to sample $y \sim \mu_p$, choose each coordinate y_i independently: $y_i = 1$ with probability p, and $y_i = 0$ with probability $1-p$. We can extend this measure more generally to $\{0,1\}^S$ (where $S \subseteq \{1,\ldots,n\}$), a measure which we refer to as $\mu_p(\{0,1\}^S)$. We sometimes identify $\mu_p(\{0,1\}^S)$ with the distribution on $\{0,1\}^n$ obtained from $\mu_p(\{0,1\}^S)$ by zero extension.

We often identify $\{0,1\}^n$ with the power set of $\{1,\ldots,n\}$. According to this view, we can sample a set from μ_p by including each element with probability p, independently. We will sometimes emphasize this view by using the notation $\mu_p(S)$ for $\mu_p(\{0,1\}^S)$.

If $f\colon \{0,1\}^n \to \mathbb{R}$, then the norm of f with respect to μ_p is $\|f\| = \sqrt{\mathbb{E}[f^2]}$, where the expectation is with respect to μ_p. We will sometimes use the notation $\|f\|_{\mu_p}$ to emphasize that the norm is taken with respect to μ_p.

We say that f, g are *ϵ-close (in L2)* if $\|f-g\|^2 \le \epsilon$.

Y-expansion Every function $f\colon \{0,1\}^n \to \mathbb{R}$ has a unique expansion as a multilinear polynomial in the input variables $y_1,\ldots,y_n$, known as the *y-expansion*. The coefficients of the monomials are called *y-coefficients*. The *degree* of f is the degree of its unique expansion. A *degree d function* is a function whose degree is *at most d*.

We will use the shorthand y_I for the monomial $\prod_{i\in I} y_i$.

The *support* of the y-expansion of f, denoted $\operatorname{supp} f$, is the hypergraph on $\{1,\ldots,n\}$ whose hyperedges are the monomials with non-zero coefficients. *Level e* of the support, denoted $\operatorname{supp}_e f$, consists of all hyperedges of uniformity (size) e.

The *restriction* of f to S, denoted $f|_S$, is the function on $\{0,1\}^S$ obtained by substituting $y_i = 0$ for all $i \notin S$.

Fourier expansion A different canonical expansion is the p-biased Fourier expansion. Every function $f\colon \{0,1\}^n \to \mathbb{R}$ has a unique expansion as a multililnear polynomial in the variables $x_i := \frac{y_i - p}{\sqrt{p(1-p)}}$:

$$f(x_1,\ldots,x_n) = \sum_{S\subseteq\{1,\ldots,n\}} \hat{f}(S)\chi_S, \text{ where } \chi_S = \prod_{i\in S} \frac{y_i - p}{\sqrt{p(1-p)}}.$$

This expansion is known as the *(p-biased) Fourier expansion*, and it depends on the value of p. The functions χ_S are known as *Fourier characters*, and the coefficients $\hat{f}(S)$ as *Fourier coefficients*. The Fourier characters form an orthonormal basis with respect to the inner product $\langle f, g\rangle = \mathbb{E}_{\mu_p}[fg]$.

The degree of a function is the maximal size of a set S such that $\hat{f}(S) \neq 0$; it coincides with the notion of degree defined using the y-expansion.

We let $f^{=d} = \sum_{|S|=d} \hat{f}(S)\chi_S$, and define $f^{<d}, f^{\le d}, f^{>d}, f^{\ge d}$ analogously.

A-valued A function f is *A-valued* if $f(y) \in A$ for all y in the domain. The distance of f from being A-valued is $\mathbb{E}[\operatorname{dist}(f,A)^2]$, where $\operatorname{dist}(f,A) = \min_{a\in A}|f-a|$. Rounding x to A, denoted $\operatorname{round}(x,A)$, means replacing x with the closest value in A (breaking ties arbitrarily).

The slice Our results extend to the *slice* or *Johnson scheme*, which consists of all points in $\{0,1\}^n$ with a specified Hamming weight. Since most of this paper does not require knowledge of the slice, we relegate its treatment to Section 3.7.

3 Main theorem

In this section we state and prove our main result, the biased Kindler–Safra theorem. The theorem has two versions: one for low degree functions which are almost A-valued, and the other for A-valued functions which are almost low degree. We state the first version as a theorem and the second one as a corollary.

Theorem 3.1. *For every integer d and finite set $A \subseteq \mathbb{R}$ there exists a constant $C \leq 1$ and a finite set $B \subseteq \mathbb{R}$ such that the following holds.*

Let $p \leq 1/2$; all expectations in the sequel are with respect to μ_p. Let $f\colon \{0,1\}^n \to \mathbb{R}$ be a degree d function, and define $\epsilon := \mathbb{E}[\operatorname{dist}(f, A)^2]$. There exists a degree d function $g\colon \{0,1\}^n \to \mathbb{R}$ satisfying the following properties.
The function g is close to f:

(a) $\|f - g\|^2 = O(\epsilon)$.

(b) $\Pr[\operatorname{round}(f, A) \neq g] = O(\epsilon)$.

It has a very specific structure:

(c) The y-coefficients of g belong to B.

(d) The support of g has branching factor $O(1/p)$: for every set T and integer $e \geq |T|$, there are at most $O(1/p^{e-|T|})$ sets at level e of the support of g which contain T. (We discuss this notion further in Section 3.2.)

(e) The support of g contains $O(\epsilon^C/p^e + 1)$ sets on level e.

It behaves like an A-valued junta:

(f) $\Pr[g \notin A] = O(\epsilon)$.

(g) If $y \sim \mu_p$ then with probability $1 - O(\epsilon)$, there are $O(1)$ monomials evaluating to 1 in the y-expansion of g.

It is almost constant:

(h) $\Pr[g \neq a] = O(\epsilon^C + p)$*, where $a \in A$ is the constant coefficient of g.*

(i) $\Pr[|g - a| \geq t] \leq e^{-t^{O(1/d)}} O(\epsilon^C + p)$ *for all $t > 0$, where a is as in (h).*

It has bounded moments:

(j) The norm of g satisfies $\mathbb{E}[g^2] = O(1)$.

(k) The variance of g satisfies $\mathbb{V}[g] = O(\epsilon^C + p)$.

(l) More generally, $\mathbb{E}[|g - a|^k] \leq k^{O(dk)}(\epsilon^C + p)$ for integer $k \geq 2$.

All big O constants depend on d and A but not on p or n.

Corollary 3.2. *For every integer d and finite set $A \subseteq \mathbb{R}$ there exists a constant $C \leq 1$ and a finite set $B \subseteq \mathbb{R}$ such that the following holds.*

Let $p \leq 1/2$; all expectations in the sequel are with respect to μ_p. Let $F\colon \{0,1\}^n \to A$ be an A-valued function, and let $\epsilon := \|F^{>d}\|^2$ be the distance of F from the closest degree d function. There exists a degree d function $g\colon \{0,1\}^n \to \mathbb{R}$ satisfying the following properties:

(a') $\|F - g\|^2 = O(\epsilon)$.

(b') $\Pr[F \neq g] = O(\epsilon)$.

(c') $\Pr[F \neq a] = O(\epsilon^C + p)$, *where* $a \in A$ *is the constant coefficient of* g.

(d') All properties of g *stated in Theorem 3.1.*

We deduce Corollary 3.2 from Theorem 3.1 in Section 3.5, where we also briefly discuss the optimal value of the constant C.

How complete is the characterization in Theorem 3.1? It turns out that a small subset of the properties proved in the theorem suffice to imply that f is close to A:

Theorem 3.3. *For every integer* d *and finite sets* $A, B \subseteq \mathbb{R}$ *the following holds.*

Let $g \colon \{0,1\}^n \to \mathbb{R}$ *be a degree* d *function whose y-coefficients belong to* B, *and let* $p \leq 1/2$. *If* g *has branching factor* $O(1/p)$ *and* $\epsilon := \Pr[g \notin A]$ *then*

$$\mathbb{E}[\operatorname{dist}(g, A)^2] = O(\epsilon).$$

Stated differently, if g is a degree d function that satisfies Items (c), (d) and (f) of Theorem 3.1, then $\mathbb{E}[\operatorname{dist}(g, A)^2] = O(\epsilon)$. We prove Theorem 3.3 in Section 3.6.

The rest of this section is organized as follows. We sketch the proof of Theorem 3.1 in Section 3.1. Several preliminary lemmas on branching factor are proved in Section 3.2, and the proof itself appears in Section 3.3 and Section 3.4. We prove and discuss Corollary 3.2 in Section 3.5, and prove Theorem 3.3 in Section 3.6. Finally, we adapt the results to the slice in Section 3.7.

3.1 Proof overview

The classical Kindler–Safra theorem describes the structure of low degree functions which are almost Boolean with respect to the measure μ_q for any fixed q. Our goal is to lift this theorem to the biased Boolean cube. The idea is to consider random restrictions of f. Choose $S \sim \mu_{2p}$, and let $f|_S$ be the restriction of f to $\{0,1\}^S$ obtained by substituting zeroes in all other coordinates. The main observation is that if we now sample a point $y \sim \mu_{1/2}(\{0,1\}^S)$, then the point y has the distribution $\mu_p(\{0,1\}^n)$. This implies that on average, the functions $f|_S$ are close (with respect to $\mu_{1/2}$) to A. Applying a generalization of the Kindler–Safra theorem to the A-valued setting, we approximate the functions $f|_S$ with juntas g_S.

Using the fact that the g_S are juntas, we can show that if we choose random S_1, S_2 with large intersection, then the two functions g_{S_1}, g_{S_2} will agree with probability $1 - O(\epsilon)$. This allows us to stitch the juntas g_S to a global function g satisfying $\Pr[g|_S = g_S] = 1 - O(\epsilon)$, using a higher-dimensional agreement theorem that we prove in Section 7.

The function g satisfies several properties by construction:

1. Item (c): g has degree d, and its y-coefficients belong to B.
2. Item (f): $\Pr[g \notin A] = O(\epsilon)$: this follows from g_S being A-valued.
3. Item (b): $\Pr[\operatorname{round}(f, A) \neq g] = O(\epsilon)$: this follows from both $\operatorname{round}(f, A)$ and g_S being A-valued.
4. Item (g): with probability $1 - O(\epsilon)$, only a constant number of monomials in the support of g evaluate to 1: this follows from g_S being a junta.

5. Item (d): g has branching factor $O(1/p)$: this follows from the fact that g is obtained from the g_S by "majority decoding".

6. Item (j): g has constant norm: this follows from g^2 also having branching factor $O(1/p)$.

Using the branching factor property, we are able to prove a large deviation bound on g, namely $\mathbb{E}[G^2] = O(\epsilon)$, where $G = \prod_{a \in A}(g - a)$. This allows us to prove Item (a), stating that $\|f - g\|^2 = O(\epsilon)$, by bounding the contribution of large values of g.

We bound the size of the support at each level of g, proving Item (e), using the following observation: if the support were too large, then too many of the corresponding monomials would evaluate to 1 on a random input, contradicting Item (g).

Item (e) implies almost directly that g is close to being constant (Item (h)) and has small variance (Item (k)). Another short argument bounds high moments of g (Item (l)) using the branching factor property. The large deviation bound, Item (i), then follows by standard techniques.

3.2 Branching factor

One of the concepts instrumental for the proof of Theorem 3.1 is that of *branching factor*.

A hypergraph has *branching factor* ρ if for any set T and any $e \geq |T|$, the number of hyperedges of uniformity e containing T is at most $\rho^{e-|T|}$.

Recall that for each function $f\colon \{0,1\}^n \to \mathbb{R}$ we define its support to be the hypergraph whose hyperedges correspond to monomials in the y-expansion with non-zero coefficient. The function f has branching factor ρ if its support has branching factor ρ.

All branching factors encountered during the proof of Theorem 3.1 are $O(1/p)$.

If a hypergraph has branching factor ρ then the number of hyperedges of uniformity e is at most ρ^e. But the branching factor property is stronger, since it is preserved (up to constants) under contraction of a constant number of vertices, that is, removing these vertices from all hyperedges:

Lemma 3.4. *Let h be a function having branching factor ρ. Suppose we substitute $y_i = 1$, for some i. The resulting function h' has branching factor 2ρ.*

Proof. Fix a set T (not containing i) and an integer e. We will bound the number of sets $S \supseteq T$ of size e in the support of h'.

If a set S is in the support of h' then one of $S, S \cup \{i\}$ must be in the support of h. In other words, the number of sets $S \supseteq T$ of size e in the support of h' is at most the number of such sets in the support of h, together with the number of sets $S' \supseteq T \cup \{i\}$ of size $e+1$ in the support of h. Since h has branching factor ρ, there are at most $\rho^{e-|T|}$ of the former and $\rho^{(e+1)-(|T|+1)}$ of the latter, for a total of at most $2\rho^{e-|T|}$. It follows that h' has branching factor 2ρ. □

Our proof of Theorem 3.1 will repeatedly use some additional properties of branching factor. The first one states that the product of two constant degree functions having branching factor ρ also has branching factor $O(\rho)$.

Lemma 3.5. *If h_1, h_2 are two degree d functions having branching factor ρ then their product $h_1 h_2$ has branching factor $O_d(\rho)$.*

Proof. Consider any set T and any integer $e \geq |T|$. Our goal is to bound the number of sets $S \supseteq T$ of size e in the support of $h_1 h_2$. We can assume that $e \leq 2d$.

If $S \supseteq T$ is in the support of $h_1 h_2$ then there must be S_1, S_2 in the supports of h_1, h_2 (respectively) such that $S_1 \cup S_2 = S$, hence it suffices to bound the number of such pairs. To do so, we generate all sets according to the following algorithm:

1. Choose a subset $T_1 \subseteq T$, which will be $S_1 \cap T$. There are $2^{|T|} = O(1)$ possible choices.
2. Choose sizes $s_1, s_2 \leq e$, which will be the sizes of S_1, S_2 (respectively). There are at most $(e+1)^2 = O(1)$ possible choices.
3. Choose a set S_1 of size s_1 in the support of h_1 containing T_1. Since h_1 has branching factor ρ, there are at most $\rho^{s_1 - |T_1|}$ possible choices.
4. Choose a subset R of S_1 of size $s_1 + s_2 - e$ satisfying $R \cap T \subseteq T_1$, which will be $S_1 \cap S_2$. There are at most $2^e = O(1)$ possible choices.
5. Choose a set S_2 of size s_2 in the support of h_2 containing $R \cup (T \setminus T_1)$. There are at most $\rho^{s_2 - (s_1 + s_2 - e + |T \setminus T_1|)} = \rho^{e - s_1 - |T \setminus T_1|}$ possible choices.

In total, the number of pairs S_1, S_2 is $O(\rho^{e - |T|})$. It follows that $h_1 h_2$ has branching factor $O(\rho)$. $\square$

We also need a variant of this argument in which we are additionally given that the levels of the y-expansion of h_1 are individually sparse.

Lemma 3.6. *If h_1, h_2 are two degree d functions and h_2 has branching factor ρ then for all $e \leq 2d$,*

$$|\operatorname{supp}_e(h_1 h_2)| = O_d\big(\max_{s \leq e} |\operatorname{supp}_s(h_1)| \rho^{e-s}\big).$$

(Recall that $\operatorname{supp}_e(h)$ consists of all sets in $\operatorname{supp} h$ of size e.)

Proof. If S is in the support of $h_1 h_2$, then there must be S_1, S_2 in the supports of h_1, h_2 (respectively) such that $S_1 \cup S_2 = S$, hence it suffices to bound the number of such pairs. We can generate them as follows:

1. Choose sizes $s_1, s_2 \leq e$, which will be the sizes of S_1, S_2 (respectively). There are $O(1)$ possible choices.
2. Choose a set S_1 of size s_1 in the support of h_1. There are $|\operatorname{supp}_{s_1}(h_1)|$ possible choices.
3. Choose a subset R of S_1 of size $s_1 + s_2 - e$, which will be $S_1 \cap S_2$. There are $O(1)$ possible choices.
4. Choose a set S_2 of size s_2 containing R. There are $\rho^{s_2 - (s_1 + s_2 - e)} = \rho^{e - s_1}$ possible choices.

Given s_1, the total number of choices is $O(|\operatorname{supp}_{s_1}(h_1)| \rho^{e - s_1})$. Summing over all s_1, we obtain the statement of the lemma. $\square$

The second property we will use is that a function with branching factor $O(1/p)$ equals its constant coefficient with constant probability.

Lemma 3.7. *If h is a degree d function which has branching factor $O(1/p)$ then*

$$\Pr_{y \sim \mu_p}[y_T = 0 \text{ for all non-empty } T \in \operatorname{supp} h] = \Omega(1).$$

Proof. For each non-empty set T in the support of h, the probability that $y_T = 0$ is $1 - p^{|T|}$, and this is an anti-monotone event. The FKG inequality states that anti-monotone events positively correlate, and so the probability that $y_T = 0$ for all non-empty sets T in the support of h is at least

$$\prod_{e=1}^{d}(1-p^e)^{|\operatorname{supp}_e h|} \geq \prod_{e=1}^{d}(1-p^e)^{O(1/p^e)} = \Omega(1),$$

using $(1-q)^{1/q} = \Omega(1)$. □

3.3 Step 1: Constructing g

We now turn to the proof of Theorem 3.1. We fix d and A, and let $f\colon \{0,1\}^n \to \mathbb{R}$ be a degree d function satisfying $\epsilon := \mathbb{E}[\operatorname{dist}(f, A)^2]$ with respect to μ_p, where $p \leq 1/2$. All asymptotic notations in this section and the next depend on d, A but not on n, p.

We start the proof by constructing the function g. For each set $S \subseteq \{1, \ldots, n\}$, recall that $f|_S$ is the restriction of f to $\{0,1\}^S$ obtained by substituting zeroes for the coordinates outside S. If we choose $S \sim \mu_{2p}(\{0,1\}^n)$ and $y \sim \mu_{1/2}(\{0,1\}^S)$ then $y \sim \mu_p(\{0,1\}^n)$, and so

$$\mathop{\mathbb{E}}_{S\sim\mu_{2p}}\Big[\mathop{\mathbb{E}}_{\mu_{1/2}}[\operatorname{dist}(f|_S, A)^2]\Big] = \mathop{\mathbb{E}}_{\mu_p}[\operatorname{dist}(f, A)^2] = \epsilon.$$

We now wish to apply the Kindler–Safra theorem on each $f|_S$, in the following form.

Theorem 3.8 (Kindler–Safra)**.** *For each integer d, finite set A, and real $\gamma > 0$, there is an integer M and a finite set B such that the following holds for all q in the range $(\gamma, 1-\gamma)$.*

If $\phi\colon \{0,1\}^n \to \mathbb{R}$ is a degree d function satisfying $\delta := \mathbb{E}_{\mu_q}[\operatorname{dist}(\phi, A)^2]$ then there exists an A-valued function ψ depending on at most M inputs, whose y-coefficients belong to B, satisfying $\mathbb{E}_{\mu_q}[(\phi - \psi)^2] = O(\delta)$.

We show how to deduce this theorem from the statement in Kindler's thesis [Kin03] in Section 5, and provide an alternative proof from scratch for the case $q = 1/2$ in Section 6.

While we state Theorem 3.8 for a range of q, in practice we will only apply it to a single value of q. However, this greater generality will be needed when adapting the proof to the slice (although we would not make this explicit), since in that case we have to choose q in such a way that qn is integral.

Applying the theorem for each $f|_S$, we obtain A-valued degree d juntas g_S, whose y-coefficients belong to B, satisfying

$$\mathop{\mathbb{E}}_{S\sim\mu_{2p}}[\|f|_S - g_S\|^2] = O(\epsilon). \tag{1}$$

We now wish to stitch together the functions g_S using an agreement theorem.

Before we can state the theorem, we need to define a distribution $\mu_{p,q}$ on triples $(S_1, S_2, T) \in (\{0,1\}^n)^3$, where $0 < q < p < 1$. To sample $(S_1, S_2, T) \sim \mu_{p,q}$, first sample $T \sim \mu_q$, and then sample each of S_1, S_2 independently by starting with T and adding each element $x \notin T$ with probability $r = \frac{p-q}{1-q}$. The choice of r guarantees that $S_1, S_2 \sim \mu_p$ (in a dependent manner), since $q + (1-q)r = p$.

We can construct the marginal distribution of (S_1, T) in the following equivalent way: sample $S_1 \sim \mu_p$, and let $T \sim \mu_{q/p}(S_1)$.

We can now state the agreement theorem, which we prove in Section 7.

Theorem 3.9 (Agreement theorem)**.** *For every two integers $d, N \geq 1$ and real $\gamma > 0$, the following holds for all q, r satisfying $\gamma \leq r/q \leq 1-\gamma$.*

Suppose that for each $S \subseteq \{1,\dots,n\}$ we are given a mapping $\phi_S \colon \binom{S}{\leq d} \to \Sigma$, where $\binom{S}{\leq d}$ is the set of all subsets of S of size at most d, and Σ is an arbitrary alphabet containing 0. Suppose furthermore that for each S there are at most N many inputs A such that $\phi_S(A) \neq 0$. Let

$$\delta := \Pr_{(S_1,S_2,T)\sim\mu_{q,r}}[\phi_{S_1}|_T \neq \phi_{S_2}|_T].$$

For every set A of size at most d, let $\psi(A)$ be a most common value of $\phi_S(T)$ among all $S \supseteq T$ (with respect to μ_q). Then

$$\Pr_{S\sim\mu_q}[\psi|_S \neq \phi_S] = O(\delta),$$

where the hidden constant depends only on d, γ, N.

We will apply this theorem with $q := 2p$ and $r := \sqrt{2}p$ to the functions g_S. Since g_S depends on at most M coordinates, its y-expansion has at most $N = \binom{M}{\leq d}$ non-zero coefficients, as needed. In order to apply the theorem, we will now bound δ.

Consider $(S_1, S_2, T) \sim \mu_{2p,\sqrt{2}p}$. Recall that we can sample S_1 and T in the following manner: $S_1 \sim \mu_{2p}$ and $T \sim \mu_{\sqrt{1/2}}(S_1)$. If we choose $y \sim \mu_{\sqrt{1/2}}(\{0,1\}^T)$ then $y \sim \mu_{1/2}(\{0,1\}^{S_1})$. Therefore (1) implies

$$\mathop{\mathbb{E}}_{(S_1,S_2,T)\sim\mu_{2p,\sqrt{2}p}}\Big[\mathop{\mathbb{E}}_{\mu_{\sqrt{1/2}}}[(f|_T - g_{S_1}|_T)^2]\Big] = O(\epsilon).$$

The same holds with S_1 and S_2 switched, and so, applying the L2 triangle inequality, we deduce

$$\mathop{\mathbb{E}}_{(S_1,S_2,T)\sim\mu_{2p,\sqrt{2}p}}\Big[\mathop{\mathbb{E}}_{\mu_{\sqrt{1/2}}}[(g_{S_1}|_T - g_{S_2}|_T)^2]\Big] = O(\epsilon).$$

The two functions $g_{S_1}|_T, g_{S_2}|_T$ are A-valued functions depending on a constant number of coordinates. Hence they are either equal, or disagree with constant probability (at least $\sqrt{1/2}^M$), in which case $\mathbb{E}_{\mu_{\sqrt{1/2}}}[(g_{S_1}|_T - g_{S_2}|_T)^2] = \Omega(1)$. This immediately implies that $\delta = O(\epsilon)$.

We now apply the agreement theorem, Theorem 3.9, obtaining a function g defined as follows:

$$g = \sum_{|T|\leq d} c_T y_T,$$

where c_T is the most popular coefficient of y_T in the functions g_S for $S \supseteq T$ (where *most popular* is defined with respect to μ_{2p}). The agreement theorem states that the function g satisfies

$$\Pr_{S\sim\mu_{2p}}[g|_S \neq g_S] = O(\epsilon). \tag{2}$$

3.4 Step 2: Structure of g

In this section we uncover the various properties of g one by one.

Simple properties — Item (c), Item (f), Item (g)

The function g has degree d by construction. Its y-coefficients belong to B by construction, thus proving Item (c). To prove Item (f), we use (2):

$$\Pr_{\mu_p}[g \notin A] = \mathop{\mathbb{E}}_{S\sim\mu_{2p}}\Big[\Pr_{\mu_{1/2}}[g|_S \notin A]\Big] \leq \Pr_{S\sim\mu_{2p}}[g|_S \neq g_S] = O(\epsilon),$$

since g_S is A-valued by construction. The proof of Item (g) is similar, this time using the fact that g_S is a junta.

Closeness in probability — Item (b)

We now prove Item (b). We use the shorthand $F = \mathrm{round}(f, A)$. The main observation is that since $F|_S$ and g_S are both A-valued, $\Pr[F|_S \neq g_S] = \Theta(\|F|_S - g_S\|^2)$. This implies that

$$\Pr[F \neq g] = \mathop{\mathbb{E}}_{S \sim \mu_{2p}} \big[\Pr_{\mu_{1/2}}[F|_S \neq g|_S]\big] \leq \Pr_{S \sim \mu_{2p}}[g|_S \neq g_S] + O\big(\mathop{\mathbb{E}}_{S \sim \mu_{2p}}[\|F|_S - g_S\|^2]\big).$$

The first term is $O(\epsilon)$ by (2). We bound the second term using the L2 triangle inequality and (1):

$$\mathop{\mathbb{E}}_{S \sim \mu_{2p}}[\|F|_S - g_S\|^2] \leq 2 \mathop{\mathbb{E}}_{S \sim \mu_{2p}}[\|F|_S - f|_S\|^2] + 2 \mathop{\mathbb{E}}_{S \sim \mu_{2p}}[\|f|_S - g_S\|^2] \leq 2\|F - f\|^2 + O(\epsilon) = O(\epsilon),$$

since $\|F - f\|^2 = \epsilon$ by definition.

Branching factor — Item (d)

Our next goal is to prove Item (d). The basic observation is that if $c_T \neq 0$ then the probability that T is in the support of g_S (among $S \supseteq T$) is at least $1/2$ (otherwise 0 would have been the most popular coefficient of y_T).

Given a set U and a number e in the range $|U| \leq e \leq d$, we have to show that the number $N_{U,e}$ of sets $T \supseteq U$ of size e such that $c_T \neq 0$ is at most $O(1/p^{e-|U|})$. Since each g_S is a junta,

$$O(1) \geq \mathop{\mathbb{E}}_{\substack{S \sim \mu_{2p} \\ S \supseteq U}}[|\operatorname{supp} g_S|] \geq \sum_{\substack{T \supseteq U \\ |T| = e}} \Pr_{\substack{S \sim \mu_{2p} \\ S \supseteq U}}[T \in \operatorname{supp} g_S] \overset{(*)}{=} (2p)^{e-|U|} \sum_{\substack{T \supseteq U \\ |T| = e}} \Pr_{\substack{S \sim \mu_{2p} \\ S \supseteq T}}[T \in \operatorname{supp} g_S]$$
$$\geq (2p)^{e-|U|} \cdot \frac{N_{U,e}}{2}.$$

A random $S \sim \mu_{2p}$ containing U contains $T \setminus U$ with probability $(2p)^{|T|-|U|}$, explaining $(*)$.

It follows that $N_{U,e} = O(1/p^{e-|U|})$, and so g has branching factor $O(1/p)$.

Norm — Item (j)

Item (j) follows almost immediately from Item (d), via Lemma 3.5. The lemma implies that g^2 also has branching factor $O(1/p)$. Every coefficient in the y-expansion of g^2 is a sum of up to $2^{O(d)}$ products of two elements from B, and in particular, each non-zero coefficient is $O(1)$. It follows that

$$\mathbb{E}[g^2] \leq \sum_{T \in \operatorname{supp} g^2} O(\Pr[y_T = 1]) = \sum_{T \in \operatorname{supp} g^2} O(p^{|T|}) \overset{(*)}{\leq} \sum_{e=0}^{2d} O(p^{-e}) \cdot O(p^e) = O(1),$$

using Item (d) for $(*)$.

The function G

Proving Item (a) is more difficult than proving Item (b), since we need to control the large deviation behavior of g: it might be that on average, f and g are very close, but sometimes g is very large, and this causes g to be far from f in L2. To show that this cannot happen, we will consider the function

$$G = \prod_{a \in A} (g - a),$$

showing that $\mathbb{E}[G^2] = O(\epsilon)$. We prove this in a sequence of claims:

1. G has branching factor $O(1/p)$. This follows from Lemma 3.5.
2. $G = 0$ with probability $1 - O(\epsilon)$. This follows immediately from Item (f).
3. The support of G contains $O(p^{-e}\epsilon)$ sets of size e. Due to the branching factor property, it suffices to show this bound for inclusion-minimal sets. To do this, we show that if T is inclusion-minimal then the probability that y_T is the only non-zero monomial (in which case $G \neq 0$) is $\Omega(p^{|T|})$, using the branching factor property. Since these events are independent and $\Pr[G \neq 0] = O(\epsilon)$, we can bound the number of such T.
4. $\mathbb{E}[G^2] = O(\epsilon)$. This follows from the support of G^2 containing $O(p^{-e}\epsilon)$ many sets of size e.

Let us now prove these claims.

Claim 3.10. *The function G has branching factor $O(1/p)$.*

Proof. Since g has branching factor $O(1/p)$ by Item (d), so do the functions $g - a$. Hence the claim immediately follows from Lemma 3.5. □

Claim 3.11. *The function G satisfies* $\Pr[G = 0] = 1 - O(\epsilon)$.

Proof. Since $G = 0$ iff $g \in A$, this follows immediately from Item (f). □

Claim 3.12. *The y-expansion of G contains $O(p^{-e}\epsilon)$ monomials at level e, for each $e \leq d$.*

Proof. We say that a set T in the support of G is *minimal* if no subset of T belongs to the support of G. We show below that the number of *minimal* sets of size e in the support of G is $O(p^{-e}\epsilon)$. We first explain why this concludes the proof, and then bound the number of minimal sets.

Let us assume that the number of minimal sets of size e' is $O(p^{-e'}\epsilon)$. Every set of size e in the support of G is a superset of some minimal set of size $e' \leq e$. For each e' there are $O(p^{-e'}\epsilon)$ minimal sets of size e'. Since G has branching factor $O(1/p)$ by Claim 3.10, each such minimal set has $O(p^{-(e-e')})$ extensions at level e of G. In total, there are $O(p^{-e'}\epsilon) \cdot O(p^{-(e-e')}) = O(p^{-e}\epsilon)$ sets of size e in the support of G for any specific value of e'. Since there are $O(1)$ possible values of e', there are in total $O(p^{-e}\epsilon)$ sets of size e in the support of G.

We now bound the number of minimal sets in the support of G. Let $y \sim \mu_p$. For each minimal set T, consider the event $\mathcal{E}_T$: "$y_T = 1$, and for each other set U in the support of G, $y_U = 0$". By construction, the events $\mathcal{E}_T$ are disjoint. Furthermore, if $\mathcal{E}_T$ occurs then $G(y) \neq 0$. We will show that $\mathcal{E}_T$ happens with probability $\Omega(p^{|T|})$, and this will allow us to bound the number of minimal sets.

The probability that $y_T = 1$ is $p^{|T|}$. Applying Lemma 3.4 successively for each $i \in T$ shows that the function G_T obtained by substituting $y_i = 1$ in G for all $i \in T$ has branching factor $O(1/p)$. For the event $\mathcal{E}_T$ to occur, we need that for each $U \neq \emptyset$ in the support of G_T, $y_U = 0$. According to Lemma 3.7, this happens with probability $\Omega(1)$. Overall, $\Pr[\mathcal{E}_T] = \Omega(p^{|T|})$.

Since the events $\mathcal{E}_T$ are independent and imply that $G \neq 0$, we have

$$\Pr[G \neq 0] \geq \sum_{T \text{ minimal}} \Pr[\mathcal{E}_T] = \sum_{T \text{ minimal}} \Omega(p^{|T|}).$$

Since $\Pr[G \neq 0] = O(\epsilon)$ we conclude that the number of minimal sets of size e is $O(p^{-e}\epsilon)$, concluding the proof. □

Lemma 3.13. *The function G satisfies $\mathbb{E}[G^2] = O(\epsilon)$.*

Proof. We start by using Lemma 3.6 to show that the support of G^2 has $O(p^{-e}\epsilon)$ sets at level e. Indeed, for each s, Claim 3.12 shows that G has $O(p^{-s}\epsilon)$ sets at level s. Since G has branching factor $O(1/p)$ by Claim 3.10, Lemma 3.6 shows that the number of sets of size e in the support of G^2 is at most

$$O\big(\max_{s \leq e} p^{-s}\epsilon \cdot p^{-(e-s)}\big) = O(p^{-e}\epsilon).$$

Each coefficient of G^2 is the sum of $O(1)$ products of $2|A|$ values from A and B (either coefficients of the y-expansion of g, or values from A arising from the definition of G). Hence each such non-zero coefficient is $O(1)$. Since the probability of $y_T = 1$ is $p^{|T|}$ (for $y \sim \mu_p$), it follows that

$$\mathbb{E}[G^2] \leq \sum_{T \in \operatorname{supp} G^2} O(p^{|T|}) \leq \sum_{e=0}^{2d|A|} O(p^{-e}\epsilon) \cdot O(p^e) = O(\epsilon).$$ □

Closeness in L2 — Item (a)

Having analyzed the function G, we can now conclude the proof of Item (a), starting with

$$\begin{aligned}\|f - g\|^2 = \mathop{\mathbb{E}}_{S \sim \mu_{2p}}\big[\mathop{\mathbb{E}}_{\mu_{1/2}}[(f|_S - g|_S)^2]\big] &\overset{(*)}{\leq} 2 \mathop{\mathbb{E}}_{S \sim \mu_{2p}}\big[\mathop{\mathbb{E}}_{\mu_{1/2}}[(f|_S - g_S)^2]\big] + 2 \mathop{\mathbb{E}}_{S \sim \mu_{2p}}\big[\mathop{\mathbb{E}}_{\mu_{1/2}}[(g|_S - g_S)^2]\big] \\ &\overset{(1)}{\leq} O(\epsilon) + 2 \mathop{\mathbb{E}}_{S \sim \mu_{2p}}\big[\mathop{\mathbb{E}}_{\mu_{1/2}}[(g|_S - g_S)^2]\big],\end{aligned}$$

using the L2 triangle inequality for $(*)$.

The idea now is to bound g^2 by G^2. When g is close to A, g^2 could be much larger than G^2. However, if $|g(y)| \geq M := \max_{a \in A} |a| + 1$ (here 1 is an arbitrary constant), then $g(y)^2 = O(G(y)^2)$. Indeed, $|g(y)| \geq M$ implies $|g(y) - a| = \Theta(|g(y)|)$, and so $G(y)^2 = \Theta(g(y)^{2|A|}) = \Omega(g(y)^2)$. In particular, since g_S is A-valued, we can bound

$$(g(y) - g_S(y))^2 \leq 2g(y)^2 + 2g_S(y)^2 = O(1 + G(y)^2).$$

In fact, we can improve this bound:

$$(g(y) - g_S(y))^2 \leq O([\![g(y) \neq g_S(y)]\!] + G(y)^2).$$

Using this, we can complete the proof:

$$\mathop{\mathbb{E}}_{S\sim\mu_{2p}}\Big[\mathop{\mathbb{E}}_{\mu_{1/2}}[(g|_S - g_S)^2]\Big] \le O\Big(\mathop{\mathbb{E}}_{S\sim\mu_{2p}}\Big[\Pr_{y\sim\mu_{1/2}(S)}[g(y)\neq g_S(y)]\Big]\Big) + O(\mathop{\mathbb{E}}_{\mu_p}[G^2]) \le$$

$$O\big(\Pr_{S\sim\mu_{2p}}[g|_S \neq g_S]\big) + O(\mathop{\mathbb{E}}_{\mu_p}[G^2]) = O(\epsilon),$$

bounding the first term using (2) and the second using Lemma 3.13.

Sparsity — Item (e)

We move on to proving Item (e). Let us fix $e \le d$. The idea is this: if the level e support of g contains too many sets, then if we choose $y \sim \mu_p$, there will be significant probability that too many monomials evaluate to 1 under y, contradicting Item (g), which states that with probability $1 - O(\epsilon)$, at most $M = O(1)$ monomials evaluate to 1 in the y-expansion of g.

Recall that $\operatorname{supp}_e g$ is level e of the support of g. Let $T_1, \ldots, T_M$ be an M-tuple of distinct sets from $\operatorname{supp}_e g$, whose union we denote by T. Let $y \sim \mu_p$, and denote by $\mathcal{E}_T$ the event "the union of monomials of g evaluating to 1 is T", where we identify a monomial with the set of variables composing it. By construction, the events $\mathcal{E}_T$ are disjoint, and by Item (g), their union has probability $O(\epsilon)$.

To complete the proof, we lower bound the probability of $\mathcal{E}_T$ and the number of such sets T, starting with the former. Given Item (d), Lemma 3.4 shows that $g_{T\leftarrow 1}$ has branching factor $O(1/p)$, and so Lemma 3.7 shows that with probability $\Omega(1)$, no non-constant monomials of $g_{T\leftarrow 1}$ evaluate to 1. Therefore

$$\Pr[\mathcal{E}_T] = \Omega(p^{|T|}) = \Omega(p^{eM}).$$

It remains to count the number of sets T which are the union of M distinct sets from $\operatorname{supp}_e g$. Let $\mathcal{T}$ be the collection of all such sets. If $|\operatorname{supp}_e g| < M$ then there are no such sets, but in this case $|\operatorname{supp}_e g| = O(1)$. Otherwise, there are $\Omega(|\operatorname{supp}_e g|^M)$ many M-tuples of distinct sets from $\operatorname{supp}_e g$. Each set T can be written as the union of M sets of size e in $O(1)$ many ways, hence

$$|\mathcal{T}| = \Omega(|\operatorname{supp}_e g|^M).$$

Since the events $\mathcal{E}_T$ are disjoint, it follows that their union $\mathcal{E}$ has probability at least $\Omega((p^e|\operatorname{supp}_e g|)^M)$. On the other hand, by Item (g), $\Pr[\mathcal{E}] = O(\epsilon)$. Therefore $|\operatorname{supp}_e g| = O(p^{-e}\epsilon^{1/M})$.

Closeness to constant — Item (h)

Item (h) follows easily from Item (e). Let us first observe that the constant coefficient $c_\emptyset$ of g belongs to A. To see this, recall that $c_\emptyset$ is the most probable value of the constant coefficient of the functions g_S from Section 3.3. Since g_S is A-valued, each such constant coefficient must belong to A.

Item (e) together with a union bound shows that the probability that $g \neq c_\emptyset$ is at most

$$\sum_{e=1}^{d} O(p^{-e}\epsilon^{1/R} + 1)\cdot p^e = O(\epsilon^{1/R} + p).$$

Variance — Item (k)

Item (k) also follows essentially from Item (e). To bound the variance of g we will bound $\|g - c_\emptyset\|^2 \ge \mathbb{V}[g]$.

The function $g - c_\emptyset$ satisfies $|\operatorname{supp}_0 g| = 0$ and $|\operatorname{supp}_s g| = O(p^{-s}\epsilon^{1/R} + 1)$ for $s \geq 1$ by Item (e). Also, $g - c_\emptyset$ has branching factor $O(1/p)$ by Item (d). Applying Lemma 3.6, we deduce that the size of the level e support of $(g - c_\emptyset)^2$ is at most

$$\max_{s\geq 1} O(p^{-s}\epsilon^{1/R} + 1) \cdot O(p^{-(e-s)}) = O(p^{-e}\epsilon^{1/R} + p^{-(e-1)}).$$

Every coefficient of $(g - c_\emptyset)^2$ is a sum of $O(1)$ products of up to two elements from B or A, and in particular has value $O(1)$. Since $\Pr[y_T = 1] = p^{|T|}$, it follows that

$$\mathbb{E}[(g - c_\emptyset)^2] \leq \sum_{T \in \operatorname{supp}(g-c_\emptyset)^2} O(p^{|T|}) \leq \sum_{e=1}^{2d} O(p^{-e}\epsilon^{1/R} + p^{-(e-1)}) \cdot p^e = O(\epsilon^{1/R} + p).$$

Moments — Item (l)

Item (l) similarly follows from Item (e) and Item (d). The idea is to estimate large moments of $g - c_\emptyset$.

We start by bounding the size of the support of $(g - c_\emptyset)^k$. Every set in the support of $(g - c_\emptyset)^k$ can be written as the union of k non-empty sets $S_1, \ldots, S_k$ in the support of g. These sets form a hypergraph on up to dk vertices, which we call the *type* of the k-tuple $S_1, \ldots, S_k$. Abstracting away the names of the vertices, we can estimate the total number of types by $(dk+1)^{dk} \leq (2dk)^{dk}$, since each of the dk "spots" in $S_1, \ldots, S_k$ can be any of the dk vertices, or empty.

Let us fix a certain type. The type fixes the intersection pattern of $S_1, \ldots, S_k$. We now estimate the number of k-tuples of sets in the support of g conforming to the type. Item (e) bounds the number of choices for S_1 by $O(\epsilon^C p^{-|S_1|} + 1)$. Given $S_1, \ldots, S_{i-1}$, the type fixes the intersection of S_i with each of $S_1, \ldots, S_{i-1}$; this translates to a constraint of the form $S_i \supseteq T$, where $T = S_i \cap (S_1 \cup \cdots \cup S_{i-1})$. The type also fixes the size of S_i. The expression for T shows that $|S_i \setminus T| = |S_i \setminus (S_1 \cup \cdots \cup S_{i-1})|$, and so Item (d) bounds the number of choices for S_i by $O(p^{-|S_i \setminus (S_1 \cup \cdots \cup S_{i-1})|})$. In total, the number of k-tuples conforming to the given type is at most $O(1)^k \cdot O(\epsilon^C p^{-c} + p^{-(c-|S_1|)})$, where

$$c = |S_1| + \sum_{i=2}^{k} |S_i \setminus (S_1 \cup \cdots \cup S_{i-1})| = |S_1 \cup \cdots \cup S_k|.$$

Since S_1 is non-empty, we can bound $p^{-(c-|S_1|)}$ by $p^{-(c-1)}$.

Let us consider a set S of size e in the support of $(g - c_\emptyset)^k$. The set S can be written as a union $S_1 \cup \cdots \cup S_k$ in at most $\binom{e}{\leq d}^k \leq (e+1)^{dk} \leq (dk+1)^{dk}$ ways. The contribution of each such k-tuple to the coefficient of y_S in $(g - c_\emptyset)^k$ is $O(1)^k$, and so the coefficient of y_S is at most $O(k)^{dk}$. Since $\Pr[y_S = 1] = p^{|S|}$ (for $y \sim \mu_p$), we can bound $\mathbb{E}[(g - c_\emptyset)^k]$ level by level:

$$\mathbb{E}[(g - c_\emptyset)^k] \leq \sum_{e=1}^{dk} O(k)^{dk} \cdot O(\epsilon^C p^{-e} + p^{1-e}) \cdot p^e \leq dk \cdot O(k)^{dk} \cdot O(\epsilon^C + p) \leq k^{O(dk)}(\epsilon^C + p).$$

Large deviation — Item (i)

Item (i) follows from Item (l) via a standard argument.

Suppose that the bound on the moments is

$$\mathbb{E}[(g - c_\emptyset)^k] \leq k^{Kdk}(\epsilon^C + p).$$

(We gave a name to the hidden constant.) Choose $L > K$ arbitrarily. Given $t \geq 0$, we take

$$k = t^{1/Ld}.$$

Let us assume for the moment that k is an even integer. Markov's inequality shows that

$$\Pr[(g - c_\emptyset)^k \geq t^k] \leq \frac{\mathbb{E}[(g - c_\emptyset)^k]}{t^k} \leq \frac{t^{(K/L)k}}{t^k}(\epsilon^C + p) = e^{-\Omega(t^{1/Ld}\log t)}(\epsilon^C + p) = e^{-t^{O(1/d)}}O(\epsilon^C + p).$$

When t is large, the same bound holds if we round k to the nearest even integer. When t is small, the bound follows from Item (h).

3.5 A-valued functions

Theorem 3.1 describes the structure of a degree d function which is close to being A-valued. In this section, we prove Corollary 3.2, which describes the structure of an A-valued function which is close to degree d.

Proof of Corollary 3.2. Let $f = F^{\leq d}$. Since F is A-valued,

$$\mathbb{E}[\operatorname{dist}(f, A)^2] \leq \mathbb{E}[(F - f)^2] = \epsilon.$$

Applying Theorem 3.1 to f, we obtain a degree d function g which satisfies:

(a″) $\|f - g\|^2 = O(\epsilon)$, by Item (a).

(b″) $\Pr[\operatorname{round}(f, A) \neq g] = O(\epsilon)$, by Item (b).

(c″) $\Pr[g \neq a] = O(\epsilon^C + p)$, where $a \in A$ is the constant coefficient of g, by Item (h).

(d″) All other properties of g listed in Theorem 3.1.

The L2 triangle inequality shows that Item (a″) implies Item (a′), which states that $\|F - g\|^2 = O(\epsilon)$.

Item (b″) implies Item (b′), but the argument is more subtle. The idea is that since F is close to f and both F and $\operatorname{round}(f, A)$ are A-valued, most of the time F equals $\operatorname{round}(f, A)$. Formalizing this idea, the L2 triangle inequality shows that

$$\|F - \operatorname{round}(f, A)\|^2 \leq 2\|F - f\|^2 + 2\|f - \operatorname{round}(f, A)\|^2 \leq 2\epsilon + \|\operatorname{dist}(f, A)\|^2 \leq 4\epsilon.$$

If $F \neq \operatorname{round}(f, A)$ then $|F - \operatorname{round}(f, A)| = \Omega(1)$, and so the foregoing implies that

$$\Pr[F \neq \operatorname{round}(f, A)] = O(\epsilon).$$

Together with Item (b″), this immediately implies Item (b′), which states that $\Pr[F \neq g] = O(\epsilon)$.

Item (b′) and Item (c″) immediately imply Item (c′), which states that $\Pr[F \neq a] = O(\epsilon^C + p)$ (recalling that $C \leq 1$); note that $\epsilon = O(1)$ since $\epsilon \leq \|F\|^2 = O(1)$. $\square$

How close can F be to a constant? Item (c′) states that if F is an A-valued function that is ϵ-close to degree d, then in fact F is $O(\epsilon^C + p)$-close to being constant. The $O(p)$ term is necessary since even if $\epsilon = 0$ (that is, if F is an A-valued degree d function), the function F need not be constant; the function y_1 is p-close to constant.

It is less clear what the optimal value of C is. Filmus [Fil16a] proves that when $d = 1$ and $A = \{0, 1\}$, then either F or $1 - F$ is $O(\epsilon)$-close to a function of the form

$$y_1 + \cdots + y_m, \text{ where } m = O(\sqrt{\epsilon}/p).$$

This function is $\Theta(pm)$-close to being constant, showing that the optimal constant in this case is $C = 1/2$.

One can generalize this construction. Let ϕ be a univariate degree d polynomial such that $\phi(s) \in A$ for $0 \le s \le K$, and consider the function

$$f = \phi(y_1 + \cdots + y_m), \text{ where } m = \epsilon^{1/(K+1)}/p.$$

For small p, the distribution of $S := y_1 + \cdots + y_m$ is roughly Poisson with mean $\epsilon^{1/(K+1)}$, and so $\Pr[S > K] = \Theta(\epsilon)$. Since the Poisson distribution decays exponentially, a short calculation shows that $\mathbb{E}[\mathrm{dist}(f, A)^2] = \Theta(\epsilon)$. On the other hand, f is $\Theta(\epsilon^{1/(K+1)})$-close to constant. This shows that $C \le 1/(K + 1)$. We conjecture that the optimal value of C is obtained by such a construction.

Surprisingly, the same kind of construction appears in work of Filmus and Ihringer [FI19] on constant degree functions on the slice, which also makes an appearance in Section 4.4. The authors consider Boolean degree d functions on the slice $\{(y_1, \ldots, y_n) \in \{0, 1\}^n : \sum_i y_i = k\}$. When k is bounded away from 0 and n, the authors show that a Boolean degree d function is necessarily a junta, thus generalizing the classical Nisan–Szegedy theorem [NS94]. However, this fails when k is very small: the function

$$\phi(y_1 + \cdots + y_{n/2})$$

is Boolean-valued if $k \le K$, but is not a junta.

Von zur Gathen and Roche [GR97] considered the minimal degree of a non-constant univariate polynomial ϕ such that $\phi(0), \ldots, \phi(K) \in \{0, 1\}$, denoting it by $K - \Gamma(K)$. They conjecture that $\Gamma(K) = O(1)$, and showed that $\Gamma(K)$ is at most the largest difference between consecutive primes in $[K]$, which was shown to be at most $O(K^{0.525})$ by Baker et al. [BHP01] (the Riemann hypothesis implies the improved bound $O(K^{1/2+\epsilon})$ for all $\epsilon > 0$). Cohen, Shpilka and Tal [CST17] considered analogous questions for $\{0, \ldots, m\}$-valued polynomials.

We are interested in the dual value $K(d)$, which is the maximal value of K for which there exists a non-constant degree d univariate polynomial ϕ satisfying $\phi(0), \ldots, \phi(K) \in \{0, 1\}$. Since a non-zero degree d polynomial has at most d roots, $K(d) \le 2d$. The result of Von zur Gathen and Roche implies that $K(d) \le d + O(d^{0.525})$, and their conjecture implies that $K(d) \le d + O(1)$. In contrast, the polynomials

$$\phi_d(s) = \sum_{e=0}^{d} (-1)^e \binom{s}{e}$$

show that $K(d) \ge 2\lfloor \frac{d}{2} \rfloor + 1$ for $d \ge 1$.

3.6 A converse

In this section we prove Theorem 3.3, which is a converse of Theorem 3.1. The proof relies on the following crucial fact: the proof of Claim 3.11, which states that $\mathbb{E}[G^2] = O(\epsilon)$ (where $G = \prod_{a \in A}(g - a)$), relies only on Item (d) and Item (f), both of which are assumptions of Theorem 3.3. Therefore $\mathbb{E}[G^2] = O(\epsilon)$ in our case as well.

The rest of the proof is reminiscent of the proof of Item (a) in Section 3.4. We showed there that for $M = \max_{a \in A} |a| + 1$, if $|g(y)| \geq M$ then $g(y)^2 = O(G(y)^2)$. A similar argument shows that if $\mathrm{dist}(g(y), A) \geq M$ then $g(y)^2 = O(G(y)^2)$, and so

$$\mathrm{dist}(g(y), A)^2 \leq M^2 [\![g(y) \notin A]\!] + O(G(y)^2).$$

This immediately implies that

$$\mathbb{E}[\mathrm{dist}(g, A)^2] \leq M^2 \Pr[g \notin A] + O(\mathbb{E}[G^2]) = O(\epsilon).$$

3.7 Adaptation to the slice

Background

So far we have only considered functions on the Boolean cube $\{0,1\}^n$ with respect to the specific measures μ_p. In this section we consider a similar domain known as the *slice* or *Johnson scheme*:

$$\binom{[n]}{k} = \left\{ (y_1, \ldots, y_n) \in \{0,1\}^n : \sum_{i=1}^{n} y_i = k \right\}.$$

We consider functions on this domain with respect to the uniform distribution ν_k, which can also be thought of as a distribution on $\{0,1\}^n$ supported on the slice.

We can extend the concept of y-expansion to the slice. While the y-expansion is no longer unique, every multilinear polynomial in the monomials y_T does represent a function on the slice. We say that a function has *degree d* if it can be expressed as a degree d polynomial.

Alternatively, Dunkl [Dun76] showed that every function on the slice has a *unique* expansion as a multilinear polynomial P of degree at most $\min(k, n-k)$ satisfying $\sum_i \partial P / \partial y_i = 0$, known as the *harmonic expansion*. The degree of the function coincides with the degree of P. For more on this point of view, consult [Fil16b, FM19].

Generally speaking, the measure ν_k behaves very similarly to the measure μ_{pn}. This has been formalized in invariance principles [FKMW18, FM19], follows from a simple coupling between the domains [Lif18], and has also manifested in an extension of the Kindler–Safra theorem to the slice, due to Keller and Klein [KK20]:

Theorem 3.14 (Keller–Klein). *For each integer d, finite set A and real $\gamma > 0$ there is an integer M and a finite set B such that the following holds for all q in the range $(\gamma, 1-\gamma)$.*

If $\phi \colon \{0,1\}^n \to \mathbb{R}$ is a degree d function on $\binom{[n]}{qn}$ satisfying $\delta := \mathbb{E}_{\nu_{qn}}[\mathrm{dist}(f, A)^2]$ then there exists an A-valued function ψ depending on at most M inputs, whose y-coefficients belong to B, satisfying $\mathbb{E}_{\nu_{qn}}[(\phi - \psi)^2] = O(\delta)$.

While Keller and Klein do not state their theorem in this way, Theorem 3.14 follows from their result in the same way that Theorem 3.8 follows from the classical Kindler–Safra theorem, as we indicate at the end of Section 5.

Theorem 3.9 also has a version on the slice. In this version, the distribution $\mu_{p,q}$ is replaced by the distribution $\nu_{k,t}$, defined as follows: to sample (S_1, S_2, T) according to $\nu_{k,t}$, let T be a random subset of $[n] = \{1, \ldots, n\}$ of size t, and let S_1, S_2 be two independent random supersets of T of size k.

Theorem 3.15 (Agreement theorem on the slice). *For every two integers d, N and real $\gamma > 0$, the following holds for all ℓ, t satisfying $\gamma < t/\ell < 1 - \gamma$.*

Suppose that for each $S \in \binom{[n]}{\ell}$ we are given a mapping $\phi_S \colon \binom{S}{\leq d} \to \Sigma$, where $\binom{S}{\leq d}$ is the set of all subsets of S of size at most d, and Σ is an arbitrary alphabet containing 0. Suppose furthermore that for each S there are at most N many inputs A such that $\phi_S(A) \neq 0$. Let

$$\delta := \Pr_{(S_1,S_2,T)\sim\nu_{\ell,t}}[\phi_{S_1}|_T \neq \phi_{S_2}|_T].$$

For every set A of size at most d, let $\psi(A)$ be a most common value of $\phi_S(T)$ among all $S \supseteq A$ (with respect to ν_ℓ). Then

$$\Pr_{S\sim\nu_\ell}[\psi|_S \neq \phi_S] = O(\delta),$$

where the hidden constant depends only on d, γ.

Sparse representations are unique

Here is how we are planning to apply Theorem 3.15: for every set $S \in \binom{[n]}{2k}$, we apply Theorem 3.14 to $f|_S$, obtaining an A-valued degree d junta g_S. We then stitch the functions g_S together to a global function g using Theorem 3.15. In order to show that the hypothesis of the theorem holds, the proof in the p-biased setting uses the following argument: if $g_{S_1}|_T \neq g_{S_2}|_T$ as coefficients of y-expansions then $g_{S_1}|_T \neq g_{S_2}|_T$ also as functions, and so $\Pr[g_{S_1}|_T \neq g_{S_2}|_T] = \Omega(1)$ since g_{S_1}, g_{S_2} are juntas.

In the case of the Boolean cube, the uniqueness of the y-expansion implies that two functions g_1, g_2 have an equal y-expansion if and only if they are equal. This is, however, no longer the case for the slice. For example, $\sum_{i=1}^n y_i - k = 0$ on the slice $\binom{[n]}{k}$ (as functions). We can rule out this particular example, for large enough n, since g_1 is not a junta; and this represents a general phenomenon, as we now spell out.

Suppose that $g_{S_1}|_T \neq g_{S_2}|_T$ as y-expansions, although $g_{S_1}|_T = g_{S_2}|_T$ as functions. Then the y-expansion $h = g_{S_1}|_T - g_{S_2}|_T$ is a *representation of zero* (that is, vanishes on its domain $\binom{[n]}{k}$) of degree d and *sparsity* (number of non-zero coefficients) at most $N = 2\binom{M}{\leq d}$. Our goal is to show that for large enough $k, n-k$ (as a function of d), this is impossible.

The first step is understanding the space of representations of zero.

Lemma 3.16. *A degree d y-expansion h is a representation of zero on the slice $\binom{[n]}{k}$, where $d \leq \min(k, n-k)$, if and only if h is the multilinearization of $(y_1 + \cdots + y_n - k)P(y_1, \ldots, y_n)$ for some polynomial P of degree at most $d-1$.*

(The *multilinearization* of a polynomial is obtained by replacing higher powers of y_i by y_i.)

Proof. Clearly every function of the form $(y_1 + \cdots + y_n - k)P$ represents zero. Furthermore, if $\deg P \leq d-1$ then the multilinearization of $(y_1 + \cdots + y_n - k)P$ has degree at most d.

Next, we claim that if $P \neq Q$ are two polynomials of degree at most $d-1$, then the multilinearizations of $(y_1 + \cdots + y_n - k)P$ and $(y_1 + \cdots + y_n - k)Q$ are different. Equivalently, if $R \neq 0$ has degree at most $d-1$ then the multilinearization of $(y_1 + \cdots + y_n - k)R$ is not the zero polynomial. Indeed, suppose that $\deg R = e < d$, and denote the multilinearization of $(y_1 + \cdots + y_n - k)R$ by S. We can think of $R^{=e}$ as a function on $\binom{[n]}{e}$ (encoding the coefficients of the various degree e monomials), and of $S^{=e+1}$ as a function on $\binom{[n]}{e+1}$. These two functions are related by the identity

$$S^{=e+1}(A) = \sum_{i\in A} R^{=e}(A \setminus \{i\}).$$

(This relation only holds for the top level of S.) We can write $S^{=e+1} = UR^{=e}$, where U is the $\binom{[n]}{e+1} \times \binom{[n]}{e}$ matrix given by $U(A,B) = 1$ if $A \supset B$.

The matrix U is the so-called *Up operator* or *raising operator* in the Boolean lattice, and is well-known to have full rank, see for example [Sta91, Theorem 2.2]. In particular, since $e+1 \le d \le \min(k, n-k) \le n/2$, this means that U is injective, and so $R \neq 0$ implies that $S \neq 0$, as required.

It follows that the space V of multilinearizations of $(y_1 + \cdots + y_n - k)P$ over all polynomials P of degree at most $d-1$ has dimension $\binom{n}{\le d-1}$.

The space of all polynomials of degree at most d has dimension $\binom{n}{\le d}$, and the space of all functions of degree at most d has dimension $\binom{n}{d}$, a classical result of Dunkl [Dun76] (see also [Fil16b, FM19]). This implies that the space of all representations of zero of degree at most d has dimension $\binom{n}{\le d-1}$, coinciding with the dimension of V. It follows that V consists of all representations of zero of degree at most d. $\square$

Armed with this folklore property, we can show that every representation of zero is somewhat dense.

Lemma 3.17. *If h is a non-zero representation of zero over $\binom{[n]}{k}$ of degree at most d, where $d \le \min(k, n-k)$, then h contains at least $\lfloor \frac{n-d+1}{d} \rfloor > \frac{n}{d} - 2$ non-zero coefficients.*

Proof. Lemma 3.16 shows that h is the multilinearization of $(y_1 + \cdots + y_n - k)P$, where $P \neq 0$ has degree at most $d-1$. Let $\deg P = e < d$, and suppose that the coefficient of y_A is non-zero, where $|A| = e$.

As in the proof of Lemma 3.16, we can think of $P^{=e}$ as a function on $\binom{[n]}{e}$, and of $h^{=e+1}$ as a function on $\binom{[n]}{e+1}$. These two functions are related by

$$h^{=e+1} = UP^{=e},$$

where U is the $\binom{[n]}{e+1} \times \binom{[n]}{e}$ inclusion matrix.

Let B be an arbitrary subset of $e+1$ elements disjoint from A (such a subset exists since $e+1 \le d \le \min(k, n-k) \le n/2$), and consider the restrictions of $h^{=e+1}$ and of $P^{=e}$ to the subsets of $A \cup B$. The restriction of U to this domain is the $\binom{[2e+1]}{e+1} \times \binom{[2e+1]}{e}$ inclusion matrix, and we have $h^{=e+1}|_{A\cup B} = U|_{A\cup B} P^{=e}|_{A\cup B}$.

Complementing the sets representing the rows, we can also think of $U|_{A\cup B}$ as the $\binom{[2e+1]}{e} \times \binom{[2e+1]}{e}$ matrix given by $U|_{A\cup B}(S,T) = 1$ if $\overline{S} \supset T$, that is, if $S \cap T = \emptyset$. This is just the adjacency matrix of an odd graph (special case of a Kneser graph), which is known to be regular (its eigenvalues are $(-1)^j \binom{e+1-j}{e-j} = (-1)^j (e+1-j) \neq 0$ for $j = 0, \ldots, e$).

By construction, $P^{=e}|_{A\cup B} \neq 0$, and so $h^{=e+1}|_{A\cup B} \neq 0$. This means that the coefficient of $y_{C(B)}$ in $h^{=e+1}$ is non-zero, for some $C(B) \subset A \cup B$ of size $e+1$. Since $|A| = e$, the set $C(B)$ must contain an element of B.

We can partition $\overline{A}$ into $\lfloor \frac{n-e}{e+1} \rfloor$ sets B_i of size $e+1$ (possibly leaving a small leftover). For each B_i, the argument above gives a monomial $y_{C(B_i)}$ with non-zero coefficient. Since $C(B_i)$ intersects B_i and the B_i are disjoint, all of these monomials are different. This completes the proof, using $e \le d-1$. $\square$

Corollary 3.18. *If g_1, g_2 are two y-expansions over $\binom{[m]}{\ell}$ of degree at most d and sparsity at most N, where $\ell, m-\ell \ge d$ and $m \ge 2d(N+1)$, then $g_1 = g_2$ as y-expansions iff $g_1 = g_2$ as functions.*

Proof of the main theorem

Replacing Theorem 3.8 with Theorem 3.14 and Theorem 3.9 with Theorem 3.15, most of the proof of Theorem 3.1 goes through on a slice $\binom{[n]}{k}$ with few changes (using Corollary 3.18 to show that the hypothesis of Theorem 3.15 indeed holds), as long as k is larger than some constant depending only on $d, |A|$; as an example, while $\Pr[y_T = 1]$ is not exactly $(k/n)^{|T|}$, for bounded $|T|$ it is $\Theta((k/n)^{|T|})$.

There is only one point in the argument which is specific to the Boolean cube, namely Lemma 3.7, which employs the FKG inequality. We can prove the corresponding result for the slice by reduction to Lemma 3.7:

Lemma 3.19. *If h is a degree d function on the slice $\binom{[n]}{pn}$ with branching factor $O(1/p)$, where $p \leq 1/2$ and pn is large enough (as a function of d and the hidden constant in $O(1/p)$), then*

$$\Pr_{y \sim \nu_{pn}}[y_T = 0 \text{ for all non-empty } T \in \operatorname{supp} h] = \Omega(1).$$

Proof. Let us denote the event in question by $\mathcal{E}$. The idea of the proof is to relate the probability of $\mathcal{E}$ under μ_{2p} to its probability under various slice distributions, and then to its probability on the specific slice $\binom{[n]}{pn}$ via monotonicity.

Lemma 3.7 shows that

$$\Pr_{\mu_{2p}}[\mathcal{E}] = \Omega(1).$$

On the other hand,

$$\Pr_{\mu_{2p}}[\mathcal{E}] = \sum_{k=0}^{n} \Pr[\text{Bin}(n, 2p) = k] \Pr_{\nu_k}[\mathcal{E}],$$

where $\text{Bin}(n, 2p)$ is the binomial distribution.

Since $\mathcal{E}$ is an anti-monotone event, the probability $\nu_k(\mathcal{E})$ is non-increasing in k, and so

$$\Pr_{\mu_{2p}}[\mathcal{E}] \leq \sum_{k=0}^{pn-1} \Pr[\text{Bin}(n, 2p) = k] \cdot 1 + \sum_{k=pn}^{n} \Pr[\text{Bin}(n, 2p) = k] \cdot \Pr_{\nu_{pn}}[\mathcal{E}] \leq \Pr[\text{Bin}(n, 2p) < pn] + \Pr_{\nu_{pn}}[\mathcal{E}].$$

Chernoff's bound shows that

$$\Pr[\text{Bin}(n, 2p) < pn] \leq e^{-pn/4},$$

and so for large enough pn, the foregoing shows that $\Pr_{\nu_{pn}}[\mathcal{E}] = \Omega(1)$. $\square$

Altogether, we conclude that the proof of Theorem 3.1 goes through for a slice $\binom{[n]}{k}$ (where $k \leq n/2$) as long as k is large enough, as a function of d and $|A|$.

Theorem 3.20. *For every integer d and finite set $A \subseteq \mathbb{R}$ there exists a constant $C \leq 1$, a finite set $B \subseteq \mathbb{R}$, and an integer k_0 such that the following holds.*

Let $f\colon \binom{[n]}{k} \to \mathbb{R}$ be a degree d function, where $k/n \leq 1/2$ and $k \geq k_0$. Define $\epsilon := \mathbb{E}[\text{dist}(f, A)^2]$, where the expectation (here and below) is taken with respect to ν_k. There exists a degree d function $g\colon \{0,1\}^n \to \mathbb{R}$ satisfying the properties listed in Theorem 3.1.

Corollary 3.21. *For every integer d and finite set $A \subseteq \mathbb{R}$ there exists a constant $C \leq 1$, a finite set $B \subseteq \mathbb{R}$, and an integer k_0 such that the following holds.*

Let $F\colon \binom{[n]}{k} \to A$ be an A-valued function, where $k/n \leq 1/2$ and $k \geq k_0$, and define $\epsilon := \|F^{>d}\|^2$. There exists a degree d function $g\colon \{0,1\}^n \to \mathbb{R}$ satisfying the properties listed in Corollary 3.2.

Here $F^{>d} = F - F^{\leq d}$, where $F^{\leq d}$ is the orthogonal projection of F to the space of degree d functions. The function $F^{>d}$ can also be obtained by removing all monomials of degree larger than d from the harmonic expansion of F.

4 Monotone functions

Recall that a Boolean function F is *monotone* if whenever two inputs $y, z \in \{0,1\}^n$ satisfy $y_i \le z_i$ for all $1 \le i \le n$, the corresponding outputs satisfy $F(y) \le F(z)$. In this section we state and prove an analog of Theorem 3.1 for monotone Boolean functions, as an illustration of the power of our proof technique.

Whereas Theorem 3.1 describes a rather complicated structure, in the monotone case we are able to approximate the function by a monotone DNF of bounded width. A *monotone DNF* is a disjunction of clauses, each consisting only of positive literals. The *width* of a DNF is the maximal number of literals in a clause.

We identify monotone DNFs with the their *support*, which is the hypergraph whose hyperedges correspond to clauses. As an example, $y_1 \vee (y_2 \wedge y_3)$ is a monotone DNF whose support is the hypergraph whose hyperedges are $\{y_1\}, \{y_2, y_3\}$.

Here is our version of Theorem 3.1 (or rather, Corollary 3.2) for monotone functions:

Theorem 4.1. *For every integer d there exists a constant $C \le 1$ such that the following holds.*

Let $p \le 1/2$; all expectations in the sequel are with respect to μ_p. Let $F\colon \{0,1\}^n \to \{0,1\}$ be a monotone Boolean function, and let $\epsilon := \|F^{>d}\|^2$ be the distance of F from the closest degree d function. There exists a monotone DNF g, of width at most d, satisfying the following properties:

(a) $\Pr[F \neq g] = O(\epsilon)$.

(b) The support of g has branching factor $O(1/p)$. (The definition of branching factor appears in Section 3.2.)

(c) The support of g contains $O(\epsilon^C/p^e + 1)$ sets on level e.

(d) If $y \sim \mu_p$ then with probability $1 - O(\epsilon)$, there are $O(1)$ clauses evaluating to 1 in g.

(e) Either $g = 1$ or $\Pr[g \neq 0] = O(\epsilon^C + p)$.

All big O constants depend on d but not on p or n.

These are the analogs of items (b),(d),(e),(g),(h) of Theorem 3.1.

The proof of Theorem 4.1 follows the same plan as that of Theorem 3.1, with minor changes. We start by proving a monotone version of the Kindler–Safra theorem in Section 4.1. We construct g in Section 4.2, and analyze its structure in Section 4.3. Finally, we generalize Theorem 4.1 to the slice (defined in Section 3.7) in Section 4.4.

It is natural to ask whether a converse to Theorem 4.1 holds. Such a result could state that if g is a monotone DNF of width at most d satisfying Item (b) and Item (c), then $\|g^{>d}\|^2$ is small. Indeed, Item (c) immediately implies Item (e), which in turn implies that $\|g^{>0}\|^2 = O(\epsilon^C + p)$.

When $d = 1$, the result of [Fil16a] implies that either $g = 1$ or g is a disjunction of at most $\max(O(\sqrt{\epsilon}/p), 1)$ variables. A monotone DNF of this form satisfies $\|g^{>1}\|^2 = O(\epsilon)$, which is tight up to a constant factor. Unfortunately, Item (b) and Item (c) do not suffice to obtain such a result in our case.

4.1 Step 0: Monotone Kindler–Safra

The DNF structure promised by Theorem 4.1 arises from a monotone version of Theorem 3.8.

Theorem 4.2. *For each integer d and real $\gamma > 0$ there is an integer M such that the following holds for all q in the range $(\gamma, 1-\gamma)$, and with respect to μ_q.*

If $f\colon \{0,1\}^n \to \{0,1\}$ is a monotone Boolean function satisfying $\delta := \|f^{>d}\|^2$ then there exists a monotone DNF g of width d, depending on at most M inputs, satisfying $\Pr[f \neq g] = O(\delta)$.

The proof relies on the following result on the structure of Boolean degree d functions, appearing in the survey of Buhrman and de Wolf [BW02].

Theorem 4.3. *Every monotone Boolean degree d function has certificate complexity at most d, and so can be represented by a monotone DNF of width d.*

We will deduce Theorem 4.2 from Theorem 3.8 by way of Theorem 4.3. The starting point is an approximation of $f^{>d}$ by a junta. If δ is small enough, then we are able to show that the junta must be a monotone degree d function, and so a monotone DNF of width d by Theorem 4.3. If δ is large, we can choose g to be a trivial DNF.

Proof of Theorem 4.2. The function $f^{\leq d}$ satisfies $\mathbb{E}[\mathrm{dist}(f^{\leq d}, \{0,1\})^2] \leq \mathbb{E}[(f^{\leq d} - f)^2] = \|f^{>d}\|^2 = \delta$, and so Theorem 3.8 (applied with $A := \{0,1\}$) states the existence of a function $g\colon \{0,1\}^n \to \{0,1\}$, depending on at most M inputs, such that $\|f^{\leq d} - g\|^2 = O(\delta)$. The L2 triangle inequality implies that $\Pr[f \neq g] = \|f - g\|^2 = O(\delta)$. For definiteness, assume that $\Pr[f \neq g] \leq C\delta$.

We wish to show that if δ is small enough, say smaller than δ_0, then g must be a monotone DNF of width d; if $\delta \geq \delta_0$ then the theorem follows trivially by taking $g = 0$.

We start by showing that if δ is small enough then g must have degree d. The main observation is

$$\|g^{>d}\|^2 \leq 2\|f^{>d}\|^2 + 2\|g^{>d} - f^{>d}\|^2 \leq 2\delta + 2\|g - f\|^2 = O(\delta),$$

using the L2 triangle inequality and the fact that $\phi \mapsto \phi^{>d}$ is an orthogonal linear projection. Since there are finitely many Boolean functions depending on M inputs, there is a constant δ_1 such that if $\|g^{>d}\|^2 < \delta_1$ then in fact $g^{>d} = 0$.[2] Let $\delta_0' = \delta_1/C$, so that $\delta < \delta_0'$ implies that $\deg g \leq d$.

A similar argument shows that if δ is small enough then g must be monotone. Without loss of generality, g depends on the first M coordinates, say $g(y_1, \ldots, y_n) = h(y_1, \ldots, y_M)$. If g isn't monotone then there exist two inputs $y, z \in \{0,1\}^M$, satisfying $y_i \leq z_i$ for $1 \leq i \leq M$, such that $h(y) = 1$ and $h(z) = 0$. For every $w \in \{0,1\}^{n-M}$, either $f(y,w) \neq h(y)$ or $f(z,w) \neq h(z)$, since f is monotone. This shows that $\Pr[f \neq g] \geq \gamma^M$. Let $\delta_0'' = \gamma^M/C$, so that $\delta < \delta_0''$ implies that g is monotone.

Define $\delta_0 = \min(\delta_0', \delta_0'')$. If $\delta < \delta_0$ then g is a monotone degree d Boolean function, and so according to Theorem 4.3, it can be written as a DNF of width d. If $\delta \geq \delta_0$ then we can trivially satisfy the theorem by taking $g = 0$. □

4.2 Step 1: Constructing g

We now turn to the proof of Theorem 4.1, fixing a value of d. Let $f\colon \{0,1\}^n \to \{0,1\}$ be a monotone Boolean function satisfying $\epsilon := \|f^{>d}\|^2$ with respect to μ_p, where $p \leq 1/2$. All asymptotic notations in this section and the next depend only on d.

[2] When $q = 1/2$, we can obtain a constructive bound on δ_1 using granularity of the Fourier coefficients: since g depends on M inputs, $\hat{g}(S)$ is an integer multiple of 2^{-M}, and so $\|g^{>d}\|^2 < 2^{-2M}$ implies $g^{>d} = 0$.

We start the proof by constructing g, closely following Section 3.3. For each set $S \subseteq \{1,\dots,n\}$, let $F|_S$ be the restriction of F to $\{0,1\}^S$, and notice that

$$\mathop{\mathbb{E}}_{S\sim\mu_{2p}}[\|(F|_S)^{>d}\|^2_{\mu_{1/2}}] \le \mathop{\mathbb{E}}_{S\sim\mu_{2p}}[\|F|_S - F^{\le d}|_S\|^2_{\mu_{1/2}}] = \|F - F^{\le d}\|^2 = \|F^{>d}\|^2 = \epsilon,$$

where the undecorated norms are with respect to μ_p; we used the fact that $\|\phi^{>d}\|$ is the distance of ϕ to the closest degree d function.

Applying Theorem 4.2 for each $F|_S$, we obtain monotone DNFs g_S of width d satisfying

$$\mathop{\mathbb{E}}_{S\sim\mu_{2p}}\Big[\Pr_{\mu_{1/2}}[F|_S \neq g_S]\Big] = O(\epsilon). \tag{3}$$

Several monotone DNFs can represent the same function. However, every monotone DNF has a unique representation in which the terms form an antichain. We call such a representation *minimal*. Without loss of generality, the monotone DNFs g_S are minimal.

We now stitch the functions g_S together using the agreement theorem, Theorem 3.9, whose formulation appears in Section 3.3. We will apply the theorem with $q := 2p$, $r := \sqrt{2}p$, and $\Sigma = \{0,1\}$ to the *supports* of the monotone DNFs g_S, which contain at most $N := \binom{M}{\le d}$ non-zero coefficients. In order to apply the agreement theorem, we need to bound the parameter

$$\delta := \Pr_{(S_1,S_2,T)\sim\mu_{2p,\sqrt{2}p}}[g_{S_1}|_T \neq g_{S_2}|_T].$$

(The definition of $\mu_{q,r}$ appears in Section 3.3.)

In this formula, $g_S|_T$ is obtained by taking the minimal monotone DNF representation of g_S, and retaining only the terms contained in T. Crucially, the resulting object is the minimal monotone DNF representation of the function obtained from g_S by zeroing out all variables outside T. Since the minimal monotone DNF representation is unique, this shows that $g_{S_1}|_T \neq g_{S_2}|_T$ as representations iff $g_{S_1}|_T \neq g_{S_2}|_T$ as functions.

Let $(S_1,S_2,T) \sim \mu_{2p,\sqrt{2}p}$. Recall that we can sample S_1 and T by first sampling $S_1 \sim \mu_{2p}$ and then sampling $T \sim \mu_{\sqrt{1/2}}(S_1)$. If we then choose $y \sim \mu_{\sqrt{1/2}}(\{0,1\}^T)$ then, given S_1, the point y has the distribution $\mu_{1/2}(\{0,1\}^{S_1})$. Therefore (3) implies that

$$\mathop{\mathbb{E}}_{(S_1,S_2,T)\sim\mu_{2p,\sqrt{2}p}}\Big[\Pr_{\mu_{\sqrt{1/2}}}[F|_T \neq g_{S_1}|_T]\Big] = O(\epsilon).$$

The same holds with S_1 and S_2 switched, and so the triangle inequality shows that

$$\mathop{\mathbb{E}}_{(S_1,S_2,T)\sim\mu_{2p,\sqrt{2}p}}\Big[\Pr_{\mu_{\sqrt{1/2}}}[g_{S_1}|_T \neq g_{S_2}|_T]\Big] = O(\epsilon).$$

The two functions $g_{S_1}|_T, g_{S_2}|_T$ both depend on at most M coordinates. Hence they are either equal, or disagree with probability at least $\sqrt{1/2}^M$. Therefore the left-hand side is lower-bounded by $\sqrt{1/2}^M\delta$, implying that $\delta = O(\epsilon)$.

Applying Theorem 3.9, we obtain a monotone DNF $\tilde{g}$ of width d satisfying

$$\Pr_{S\sim\mu_{2p}}[\tilde{g}|_S \neq g_S] = O(\epsilon),$$

where "$\tilde{g}|_S \neq g_S$" is a condition on monotone DNF representations.

Finally, g is the minimal monotone DNF representation of the function represented by $\tilde{g}$. Equivalently, g contains all inclusion-minimal terms of $\tilde{g}$. If $g|_S \neq \tilde{g}|_S$ then $\tilde{g}|_S$ is not minimal, and in particular $\tilde{g}|_S \neq g_S$ (as representations). Therefore $g|_S \neq g_S$ implies $\tilde{g}|_S \neq g_S$, and so

$$\Pr_{S \sim \mu_{2p}}[g|_S \neq g_S] = O(\epsilon). \tag{4}$$

Furthermore, if a clause C appears in g then it also appears in $\tilde{g}$, and so it appears with probability at least $1/2$ among the functions g_S for $S \supseteq C$.

4.3 Step 2: Structure of g

We now analyze the structure of g, closely following the reasoning in Section 3.4.

Item (a) follows from a simple calculation:

$$\Pr[F \neq g] = \mathop{\mathbb{E}}_{S \sim \mu_{2p}}\left[\Pr_{\mu_{1/2}}[F|_S \neq g|_S]\right] \leq \mathop{\mathbb{E}}_{S \sim \mu_{2p}}\left[\Pr_{\mu_{1/2}}[F|_S \neq g_S]\right] + \Pr_{S \sim \mu_{2p}}[g_S \neq g|_S] = O(\epsilon),$$

combining (3) and (4).

Item (d) follows directly from (4), since g_S is a junta.

Item (b) and Item (c) follow from arguments virtually identical to the proofs of their counterparts, Item (d) and Item (e), in Section 3.4. In the proof of Item (c), we crucially rely on the minimality of g, which implies that if $\mathcal{E}_T$ holds then $g|_S = g|_T$ depends on all coordinates in T.

Item (e) follows from a simple union bound given Item (c), as in Section 3.4.

4.4 Adaptation to the slice

We showed how to generalize Theorem 3.1 to the slice in Section 3.7. The same reasoning allows us to generalize Theorem 4.1 to the slice, once we have generalized Theorem 4.3 to this setting. This will require the following result of Filmus and Ihringer [FI19].

Theorem 4.4 (Filmus–Ihringer). *There exist constants C_1, C_2 such that if $C_1^d \leq k \leq n - C_1^d$ and $F \colon \binom{[n]}{k} \to \{0,1\}$ has degree d, then F depends on $K \leq C_2 2^d$ coordinates, in the sense that there are indices $i_1 < \cdots < i_K$ and a function $f \colon \{0,1\}^K \to \{0,1\}$ such that $F(y_1, \ldots, y_n) = f(y_{i_1}, \ldots, y_{i_K})$.*

Given this result, we can deduce a version of Theorem 4.3 on the slice.

Theorem 4.5. *For every integer d there is an integer k_0 such that the following holds.*

If $F \colon \binom{[n]}{k} \to \{0,1\}$ is a Boolean degree d function, where $k_0 \leq k \leq n/2$, then F can be represented by a DNF of width d.

Proof. Let $k_0 = \max(C_1^d, C_2 2^d)$. Theorem 4.4 shows that F depends on $K \leq C_2 2^d$ coordinates, without loss of generality the first K ones. Thus there is a function $f \colon \{0,1\}^K \to \{0,1\}$ such that $F(y_1, \ldots, y_n) = f(y_1, \ldots, y_k)$. Below we show that f has degree d. Theorem 4.3 shows that f, and so F, can be represented by a DNF of width d.

It remains to show that f has degree d. Suppose that this is not the case. Then $\hat{f}(S) \neq 0$ for some set $S = \{i_1, \ldots, i_s\}$, where $s > d$. If $i \notin S$ then $\frac{\partial}{\partial y_i} \chi_S = 0$, and if $i \in S$ then $\frac{\partial}{\partial y_i} \chi_S = \sqrt{p(1-p)}^{-1} \chi_{S \setminus \{i\}}$. It follows that $\left(\frac{\partial^s}{\partial y_{i_1} \cdots \partial y_{i_s}} f\right)(0) \neq 0$, where we identify f with its Fourier expansion. Since the Fourier expansion is multilinear, we can also interpret the operator $\frac{\partial}{\partial y_i}$ discretely: $\frac{\partial}{\partial y_1} f = f(1,\cdot) - f(0,\cdot)$. It follows that

$$\sum_{T \subseteq S} (-1)^{|T|} f(\mathbf{1}_T) \neq 0,$$

where $\mathbf{1}_T$ is the indicator vector of T.

Consider now a point y on the slice such that $y|_{\{1,\dots,k\}} = \mathbf{1}_S$ and $y_{K+1} = \cdots = y_{K+s} = 0$ (such a point exists since $k_0 \geq K$). For $T \subseteq S$, let $\pi(T)$ be the permutation on $\{1,\dots,n\}$ which switches i and $K+i$ for all $i \in T$. Then

$$\sum_{T \subseteq S} (-1)^{|T|} F^{\pi(T)}(y) \neq 0, \tag{5}$$

where $F^{\pi(T)}$ is obtained from F by applying $\pi(T)$ on the input.

Since F has degree d, its harmonic expansion (see Section 3.7) has degree at most d. This implies [FM19] that F is a linear combination of functions of the form $\phi = (x_{a_1} - x_{b_1}) \cdots (x_{a_e} - x_{b_e})$, where $e \leq d$. The reader can check that $\phi - \phi^{(a\ b)}$ is another function of this form, in which $x_a - x_b$ is one of the factors.[3] It follows that

$$\sum_{T \subseteq S} (-1)^{|T|} \phi^{\pi(T)}$$

is always a multiple of $(x_{i_1} - x_{K+i_1}) \cdots (x_{i_s} - x_{K+i_s})$. Since $\deg \phi < s$, it follows that $\phi = 0$. Since F is a linear combination of such functions, it follows that

$$\sum_{T \subseteq S} (-1)^{|T|} F = 0,$$

contradicting (5). This contradiction shows that f has degree d. $\square$

We also have to adapt Corollary 3.18 to our setting.

Lemma 4.6. *Suppose that g_1, g_2 are two minimal monotone DNFs of width at most d and sparsity at most N. If $\ell \geq d$ and $m - \ell \geq dN$ then $g_1 = g_2$ as DNFs iff $g_1 = g_2$ as functions over $\binom{[m]}{\ell}$.*

We comment that without the assumption of sparsity, there are counterexamples such as $y_1 \vee \cdots \vee y_{m-\ell+1}$, which equals 1 over $\binom{[m]}{\ell}$.

Proof. It suffices to show that if $g_1 \neq g_2$ as DNFs then $g_1(y) \neq g_2(y)$ for some $y \in \binom{[m]}{\ell}$. Without loss of generality, there is a term T appearing in g_1 but not in g_2. Since $\ell \geq d$ and $m - \ell \geq dN$, we can find $y \in \binom{[m]}{\ell}$ such that $y_T = 1$ and $y_i = 0$ for any variable $i \notin T$ mentioned in g_2. By construction $g_1(y) = 1$, and since g_2 is minimal, $g_2(y) = 0$. $\square$

Substituting Theorem 4.5 for Theorem 4.3 and Lemma 4.6 for Corollary 3.18, and adjusting the proof as in Section 3.7, we obtain the following version of Theorem 4.1 for the slice.

Theorem 4.7. *For every integer d there exists a constant $C \leq 1$ and an integer k_0 such that the following holds.*

Let $F\colon \binom{[n]}{k} \to \{0,1\}$ be a Boolean function, where $k/n \leq 1/2$ and $k \geq k_0$, and define $\epsilon := \|F^{>d}\|^2$. There exists a monotone DNF g of width at most d satisfying the properties listed in Theorem 4.1.

[3] Applying this operation to $x_a - x_b$, we get $2(x_a - x_b)$; to $x_a - x_i$, we get $x_a - x_b$; and to $(x_a - x_i)(x_b - x_j)$, we get $(x_a - x_b)(x_i - x_j)$.

5 A-valued Kindler–Safra theorem

The classical Kindler–Safra theorem is formulated for Boolean-valued functions. In this section we show how to reduce the A-valued version of this theorem to its Boolean-valued version, thus proving Theorem 3.8.

Nisan–Szegedy theorem

As a warm-up, we commence with the Nisan–Szegedy theorem [NS94], which is the zero-error version of the Kindler–Safra theorem. Here is the classical version:

Theorem 5.1. *For every integer $d \geq 0$ there exists an integer N_d such that every Boolean degree d function depends on at most N_d coordinates.*

(It is known that $N_d = \Theta(2^d)$, see [CHS20, Wel19].)

We prove the following A-valued version.

Theorem 5.2. *Every A-valued degree d function depends on at most $|A|N_{|A|d}$ coordinates, where N_d is the parameter from Theorem 5.1.*

The idea is very simple: if f is A-valued, then we can express it as a weighted sum of Boolean-valued functions:

$$f = \sum_{a \in A} a f_a, \text{ where } f_a = \prod_{\substack{b \in A \\ b \neq a}} \frac{f-b}{a-b}. \tag{6}$$

Using this expression, we can deduce the A-valued Nisan–Szegedy theorem immediately from its Boolean-valued version.

Proof of Theorem 5.2. Each of the functions f_a in (6) is a Boolean function of degree at most $|A|d$, hence depends on at most $N_{|A|d}$ coordinates. In total, f depends on at most $|A|N_{|A|d}$ coordinates. □

We comment that $|A|$ can be improved to $|A| - 1$ in both positions, by replacing one of the summands in (6) by a constant term, and by using $\deg f_a \leq (|A| - 1)d$.

Hypercontractivity

Before discussing the A-valued Kindler–Safra theorem, let us briefly survey hypercontractivity and some of its consequences; for more information, we refer the reader to O'Donnell's monograph [O'D14].

The context is functions $f \colon \{0,1\}^n \to \mathbb{R}$ analyzed with respect to the measure μ_p. The noise operator T_ρ maps a function f to the function $\sum_d \rho^d f^{=d}$, where $f^{=d}$ is defined according to the p-biased Fourier expansion. Taking $\lambda = \min(p, 1-p)$, for any $q \geq 2$ we have the following *hypercontractive inequality*:

$$\|T_\rho f\|_q \leq \|f\|_2, \text{ where } \rho = \lambda^{1/2-1/q}.$$

Applying this to $T_{\rho^{-1}} f$, we deduce that if f has degree d then

$$\|f\|_q \leq \lambda^{-(1/2-1/q)d} \|f\|_2.$$

Stated differently,

$$\mathbb{E}[|f|^q] \leq \lambda^{-(q/2-1)d}\, \mathbb{E}[f^2]^{q/2}.$$

If $\lambda = \Omega(1)$ and q, d are constant, then this states that $\mathbb{E}[|f|^q] = O(\mathbb{E}[f^2]^{q/2})$.

Kindler–Safra theorem

Generalizing the Kindler–Safra theorem to the A-valued setting is somewhat more complicated. However, first we have to deduce a Boolean-valued version of the theorem, whose usual statement differs from Theorem 3.8. Here is the Kindler–Safra theorem as it essentially appears in Guy Kindler's thesis [Kin03, Theorem 12.3]:

Theorem 5.3 (Kindler–Safra)**.** *Let $f\colon \{0,1\}^n \to \{-1,1\}$ satisfy $\epsilon := \|f^{>d}\|^2$ with respect to μ_p. If $\epsilon \leq c_{p,d}$ then f is $K_{p,d}\epsilon$-close (with respect to μ_p) to a ± 1-valued function depending on $J_{p,d}$ coordinates, where $c_{p,d}, K_{p,d}, J_{p,d} > 0$ are continuous functions of p.*

Let us now state and prove a Boolean-valued version of the Kindler–Safra theorem along the lines of Theorem 3.8.

Corollary 5.4. *Fix an integer $d > 0$ and a parameter $\gamma > 0$. There exists an integer J_d and a finite set B_d such that the following holds for all $p \in (\gamma, 1-\gamma)$.*

If $\phi\colon \{0,1\}^n \to \mathbb{R}$ is a degree d function satisfying $\epsilon := \mathbb{E}_{\mu_p}[\mathrm{dist}(\phi, \{0,1\})^2]$ then there exists a $0,1$-valued function ψ depending on at most J_d coordinates, whose y-coefficients belong to B_d, satisfying $\mathbb{E}[(\phi-\psi)^2] = O(\epsilon)$.

Proof. Notice first that the assumption $\gamma < p < 1-\gamma$ implies that $c_{p,d} = \Omega(1)$, $K_{p,d} = O(1)$ and $J_{p,d} \leq J_d$, for some constant J_d.

If $\epsilon > c_{p,d}$ then the theorem is satisfied by $\psi = 0$, since

$$\mathbb{E}[\phi^2] \leq 2\,\mathbb{E}[\mathrm{round}(\phi, \{0,1\})^2] + 2\,\mathbb{E}[(\phi - \mathrm{round}(\phi, \{0,1\}))^2] \leq 2 + 2\epsilon = O(\epsilon),$$

by the L2 triangle inequality. We can therefore assume that $\epsilon \leq c_{p,d}$.

Let $f = \mathrm{round}(\phi, \{0,1\})$. Since ϕ has degree d, $\|f^{>d}\|^2 \leq \|f-\phi\|^2 = \epsilon$ (this is since $f^{>d} = f - f^{\leq d}$, and $f^{\leq d}$ is the orthogonal projection of f to the space of degree d functions). Hence we can apply Theorem 5.3 (converting between $\{0,1\}$ and $\{-1,1\}$), deducing that f is $O(\epsilon)$-close to a $0,1$-valued function ψ depending on J_d coordinates.

Since ψ is a $0,1$-valued function depending on J_d coordinates, up to the choice of coordinates there are only finitely many choices for ψ. Hence the set B_d of all y-coefficients appearing in them is finite.

Finally, the L2 triangle inequality shows that

$$\mathbb{E}[(\phi-\psi)^2] \leq 2\,\mathbb{E}[(f-\phi)^2] + 2\,\mathbb{E}[(f-\psi)^2] = O(\epsilon). \qquad \square$$

In the rest of this section, we prove Theorem 3.8 (using f for ϕ and g for ψ), by reduction to Corollary 5.4. To avoid trivialities, we assume that ϵ is small, say $\epsilon \leq 1$. Otherwise, the theorem follows simply by taking $g = a$ for an arbitrary $a \in A$, as in the proof of Corollary 5.4:

$$\mathbb{E}[(f-a)^2] \leq 2\,\mathbb{E}[(\mathrm{round}(f,A)-a)^2] + 2\,\mathbb{E}[(f-\mathrm{round}(f,A))^2] = O(1+\epsilon) = O(\epsilon).$$

The idea is still to use (6) (which no longer holds with equality), but the details will be more complicated. Recall that for $a \in A$, we defined

$$f_a = \prod_{b \neq a} \frac{f-b}{a-b}.$$

In view of applying Corollary 5.4 to f_a, we need to bound $\mathbb{E}[\mathrm{dist}(f_a, \{0,1\})^2]$. We will do so in terms of $\delta = f - \mathrm{round}(f,A)$, using the notation $F = \mathrm{round}(f,A)$ for succinctness (so $f = F + \delta$).

When $F(y) = a$, we expect $f_a(y)$ to be close to 1, and when $F(y) \neq a$, we expect it to be close to 0. Let us start with the latter case. If $F(y) \neq a$ then

$$\operatorname{dist}(f_a(y), \{0,1\}) \leq |f_a(y) - 0| = \frac{|\delta(y)|}{|a - F(y)|} \prod_{b \neq a, F(y)} \frac{|F(y) + \delta(y) - b|}{|a - b|} = O(\delta(y) + \delta(y)^{|A|}).$$

(A tight estimate is $O(\delta(y))$ when $\delta(y)$ is small, and $O(\delta(y)^{|A|-1})$ when $\delta(y)$ is large.)

Similarly, when $F(y) = a$ then

$$\operatorname{dist}(f_a(y), \{0,1\}) \leq |f_a(y) - 1| = \left| \prod_{b \neq a} \left| 1 + \frac{\delta(y)}{a - b} \right| - 1 \right| = O(\delta(y) + \delta(y)^{|A|}).$$

Putting both bounds together, we obtain using the L2 triangle inequality that

$$\mathbb{E}[\operatorname{dist}(f_a, \{0,1\})^2] \leq O(\mathbb{E}[\delta^2]) + O(\mathbb{E}[\delta^{2|A|} [\![|\delta| \geq 1]\!]]).$$

Since $\mathbb{E}[(F - f)^2] = \epsilon$ by definition, the first summand is $O(\epsilon)$. We won't be so lucky with the second summand, that we bound using the Cauchy–Schwarz inequality by

$$O\left(\sqrt{\mathbb{E}[\delta^{4|A|}]} \sqrt{\Pr[|\delta| \geq 1]}\right).$$

Markov's inequality shows that $\Pr[|\delta| \geq 1] = \Pr[\delta^2 \geq 1] \leq \epsilon$. Regarding the other factor, clearly $\delta^{4|A|} \leq (2|f|)^{4|A|} + (2|F|)^{4|A|}$. Since $|F| = O(1)$, so far we have shown that

$$\mathbb{E}[\operatorname{dist}(f_a, \{0,1\})^2] = O(\epsilon) + O(\sqrt{\epsilon}) \cdot \sqrt{1 + \mathbb{E}[f^{4|A|}]},$$

the second term swallowing the first (recall we assumed ϵ is small). Since f has degree d and $p \in (\gamma, 1 - \gamma)$, hypercontractivity implies that $\mathbb{E}[f^{4|A|}] = O(\mathbb{E}[f^2]^{2|A|})$. The L2 triangle inequality shows that $\mathbb{E}[f^2] \leq 2\|f - F\|^2 + 2\|F\|^2 \leq 2\epsilon + O(1) = O(1)$, and so we can finally conclude that

$$\mathbb{E}[\operatorname{dist}(f_a, \{0,1\})^2] = O(\sqrt{\epsilon}).$$

The reader might be alarmed about the appearance of $\sqrt{\epsilon}$ rather than ϵ. Indeed, the argument below will construct an A-valued junta which is only $O(\sqrt{\epsilon})$-close to f. Fortunately, using hypercontractivity we will be able to upgrade this to $O(\epsilon)$-closeness (a similar argument appears in [KK20]).

The next step is clear: applying the Boolean-valued Kindler–Safra theorem, Corollary 5.4, we obtain for each $a \in A$ a $0,1$-valued junta g_a, depending on J_d coordinates, such that $\|f_a - g_a\|^2 = O(\sqrt{\epsilon})$. This suggests taking for g the following function:

$$g = \sum_{a \in A} a g_a.$$

Indeed, repeated application of the L2 triangle inequality shows that

$$\|f - g\|^2 = O\left(\sum_a a^2 \|f_a - g_a\|^2\right) = O(\sqrt{\epsilon}).$$

In order to complete the proof, we will deduce now that in fact $\|f-g\|^2 = O(\epsilon)$, an argument which we call *bootstrapping*.

The idea is to combine together the following two properties of $f-g$:

1. $f-g$ is ϵ-close to $A-A := \{a-b : a,b \in A\}$, since $\mathbb{E}[\operatorname{dist}(f-g, A-A)^2] \le \mathbb{E}[\operatorname{dist}(f,A)^2] = \epsilon$.
2. $\|f-g\|^2 = O(\sqrt{\epsilon})$, as we have seen above.

Since we will also use this argument again in Section 6, let us abstract it as a lemma.

Lemma 5.5 (Bootstrapping)**.** *Suppose that h is a degree d function which satisfies $\epsilon := \mathbb{E}[\operatorname{dist}(h,B)^2]$, for some finite set B. If $\|h\|^2 = O(\epsilon^{1/s})$ for some integer $s \ge 1$, then in fact $\|h\|^2 \le \epsilon + O(\epsilon^t)$ for every integer $t \ge 1$.*

The hidden constant in the conclusion depends on d, B, s, t and on the hidden constant in the premise.

Applying this argument with $h := f-g$, $d := \max(d, J_d)$, $B := A-A$, $s := 2$, and $t := 1$, we conclude that $\|f-g\|^2 = O(\epsilon)$. (As an aside, applying it with $t := 2$, we can conclude the improved bound $\|f-g\|^2 \le \epsilon + O(\epsilon^2)$.)

Proof of Lemma 5.5. Let $B' = B \cup \{0\}$.

Intuitively, since h has small norm, the only way in which it can be close to B' is if it is close most of the time to 0. This is good for us since we know that h is ϵ-close to B' rather than just $O(\epsilon^{1/s})$-close to 0. We use hypercontractivity to bound the contribution from points at which h is not closest to 0.

To flesh out our belief that h is most of the time close to 0, we use

$$\mathbb{E}[h^2] \le \mathbb{E}[h^2 [\![\operatorname{round}(h,B') \ne 0]\!]] + \mathbb{E}[\operatorname{dist}(h,B')^2].$$

The second term is at most ϵ. As for the first term, if $\operatorname{round}(h,B') \ne 0$ then $|h| = \Omega(1)$ and so $|h|^2 = O(|h|^{2st})$. This allows us to bound the first term by $O(\mathbb{E}[h^{2st}])$. Since h has degree at most d, hypercontractivity shows that

$$\mathbb{E}[h^{2st}] = O(\mathbb{E}[h^2]^{st}) = O(\epsilon^t).$$

In total, we have shown that $\mathbb{E}[h^2] \le \epsilon + O(\epsilon^t)$. □

Adaptation to the slice

The proof of Theorem 3.14 is very similar to the proof of Theorem 3.8.

The definition of the noise operator is slightly different, and the optimal value of ρ is different, but the end result $\mathbb{E}[|f|^q] = O(\mathbb{E}[f^2]^{q/2})$ still holds; see [KK20, §3].

The Kindler–Safra theorem for the slice due to Keller and Klein [KK20, Theorem 1.4] has essentially the same form as Theorem 5.3. Instead of the promise that the approximating function g depends on $J_{p,d}$ coordinates, the result of Keller and Klein promises that the approximating function has degree d. In view of Theorem 4.4, this shows that g is a junta as long as $C_1^d \le qn \le n - C_1^d$. This condition can only fail if $n < C_1^d / \min(q, 1-q)$, in which case g is trivially a junta.

The rest of the proof goes through without any changes.

6 New proof of the Kindler–Safra theorem

In this section we present a new proof of the Kindler–Safra theorem, which proceeds by induction on the degree. For simplicity, we will only consider the unbiased measure $\mu_{1/2}$.

Theorem 6.1. *For every integer d and finite set $A \subseteq \mathbb{R}$ there exists a finite set $B \subseteq \mathbb{R}$ such that the following holds, with respect to the uniform measure $\mu_{1/2}$ on $\{\pm 1\}^n$.*

If $f \colon \{\pm 1\}^n \to \mathbb{R}$ is a degree d function satisfying $\epsilon := \mathbb{E}[\operatorname{dist}(f, A)^2]$ then there exists a degree d A-valued function g, whose Fourier coefficients belong to B, satisfying $\|f - g\|^2 = O(\epsilon)$.

After switching the domain to $\{\pm 1\}^n$, the Fourier expansion takes the particularly simple form

$$f = \sum_{S \subseteq \{1,\ldots,n\}} \hat{f}(S) x_S, \text{ where } x_S = \prod_{i \in S} x_i.$$

This formula is the same as the one for the y-expansion, the only difference being that $x_i \in \{\pm 1\}$, whereas in the y-expansion $y_i \in \{0, 1\}$.

The Fourier characters x_S form an orthonormal basis, and so $\|f\|^2 = \sum_S \hat{f}(S)^2$.

During the proof, we will assume in several places that ϵ is small enough, and so $\|f\|^2 = O(1)$. If $\epsilon = \Omega(1)$ then for any $a \in A$, the L2 triangle inequality shows that

$$\|f - a\|^2 = O(\|f - \operatorname{round}(f, A)\|^2 + \|\operatorname{round}(f, A) - a\|^2) = O(\epsilon + 1) = O(\epsilon),$$

trivializing Theorem 6.1. Conversely, if $\epsilon = O(1)$ then the L2 triangle inequality shows that

$$\|f\|^2 = O(\|f - \operatorname{round}(f, A)\|^2 + \|\operatorname{round}(f, A)\|^2) = O(\epsilon + 1) = O(1).$$

Before saying more about the proof, let us show that Theorem 6.1 implies other formulations of the Kindler–Safra theorem, starting with Theorem 3.8. The only difference between Theorem 3.8 and Theorem 6.1 is that the former talks about the y-expansion while the latter talks about the Fourier expansion.

Proof of Theorem 3.8 for $\mu_{1/2}$. If f satisfies the premise of Theorem 3.8 then it also satisfies the (identical) premise of Theorem 6.1, and so we can apply the latter theorem, obtaining a function g satisfying $\|f - g\|^2 = O(\epsilon)$.

If ϵ is large then we can take $g = a$, and otherwise $\|f\|^2 = O(1)$. The L2 triangle inequality shows that also $\|g\|^2 = O(1)$. Since the Fourier coefficients of g are B-valued, g must have $O(1)$ non-zero Fourier coefficients, and so it depends on $O(1)$ coordinates. It follows that up to a choice of coordinates there are only finitely many functions g, and so the set of y-coefficients in all of them is finite. □

Another form of the Kindler–Safra theorem is the one appearing in Guy Kindler's thesis [Kin03], Theorem 5.3, which we paraphrase as follows:

Corollary 6.2. *For every integer $d > 0$ and finite set $A \subseteq \mathbb{R}$ there exists a constant $J_{d,A}$ such that the following holds.*

Let $F \colon \{\pm 1\}^n \to A$ satisfy $\epsilon := \|F^{>d}\|^2$. Then F is $O(\epsilon)$ close to an A-valued function depending on $J_{d,A}$ coordinates.

Proof. Let $f = F^{\leq d}$, so that $\mathbb{E}[\operatorname{dist}(f, A)^2] \leq \mathbb{E}[(f - F)^2] = \epsilon$. Theorem 3.8 shows that $\|f - g\|^2 = O(\epsilon)$ for some A-valued function g depending on at most $J_{d,A}$ coordinates. Applying the L2 triangle inequality concludes the proof. □

The inductive proof of Theorem 6.1 has two base cases: $d = 0$ and $d = 1$. The case $d = 0$ is trivial: take $g = \operatorname{round}(f, A)$. The case $d = 1$ is an A-valued analog of the Friedgut–Kalai–Naor (FKN) theorem [FKN02], whose proof in Kindler's thesis [Kin03, Chapter 15] we closely follow.

The proof for general d relies on both the case $d-1$ and the base case $d=1$. Crucially, we apply the case $d-1$ for a set other than A. Therefore the fact that Theorem 6.1 is stated in the A-valued setting is necessary for the proof.

6.1 A-valued FKN theorem

In this section we prove the Friedgut–Kalai–Naor theorem [FKN02], which is the base case $d=1$ of Theorem 6.1. Our proof is a simplification of the one in Kindler's thesis [Kin03, Chapter 15], replacing Kindler's use of the Azuma inequality with a simple application of hypercontractivity.

We are given a degree 1 function $f\colon \{\pm 1\}^n \to \mathbb{R}$, with Fourier expansion

$$f = \hat{f}(\emptyset) + \sum_{i=1}^{n} \hat{f}(\{i\})x_i.$$

In addition, $\mathbb{E}[\operatorname{dist}(f,A)^2] = \epsilon$. Since

$$\hat{f}(\{i\}) = \frac{f|_{x_i=1} - f|_{x_i=-1}}{2},$$

we easily obtain that the non-constant Fourier coefficients are individually quantized: since $\Pr[x_i = \pm 1] = 1/2$, clearly $\mathbb{E}[\operatorname{dist}(f|_{x_i=\pm 1}, A)^2] \le 2\epsilon$, and so the L2 triangle inequality shows that

$$\operatorname{dist}(\hat{f}(\{i\}),\Delta)^2 = O(\epsilon), \text{ where } \Delta = \{\tfrac{a-b}{2} : a,b \in A\}.$$

In order to prove the base case of Theorem 6.1, we need to show that not only are the Fourier coefficients individually quantized, but moreover, they are quantized in aggregate, that is, $\sum_{i=1}^n \operatorname{dist}(\hat{f}(\{i\},\Delta)^2 = O(\epsilon)$; and we have to show that $\hat{f}(\emptyset)$ is also quantized.

Our first step is isolating the "junta" coefficients, which are the coefficients *not* closest to 0. Every such coefficient has magnitude $\Omega(1)$, and so there are $O(1)$ such coefficients, since $\|f\|^2 = O(1)$. For definiteness, let us assume that the these coefficients are $\hat{f}(\{1\}),\ldots,\hat{f}(\{m\})$, where $m = O(1)$.

The main thrust of the proof is showing that for all $k \le n$,

$$\sum_{i=m+1}^{k} \hat{f}(\{i\})^2 \le 2\epsilon. \tag{7}$$

(In fact, this will only hold assuming ϵ is small enough.) We will prove this by induction, using Lemma 5.5.

The base case $k=m$ trivially holds, so suppose that (7) holds for some value of k. Since $\hat{f}(\{k+1\})^2 = O(\epsilon)$ by assumption, we can bound

$$\sum_{i=m+1}^{k+1} \hat{f}(\{i\})^2 = O(\epsilon).$$

Since $\mathbb{E}[\operatorname{dist}(f,A)^2] = \epsilon$, we can substitute values for all coordinates except $x_{m+1},\ldots,x_{k+1}$ so that the obtained function ϕ satisfies $\mathbb{E}[\operatorname{dist}(\phi,A)^2] \le \epsilon$. The remaining non-empty Fourier coefficients of ϕ are the same as that of f, and so $\mathbb{V}[\phi] = O(\epsilon)$. We want to show that in fact $\mathbb{V}[\phi] \le 2\epsilon$, thus proving (7) for $k+1$.

We would like to apply Lemma 5.5, but there is a crucial difference: there the information is on the L2 norm of a function, and here the information is on its variance. This suggests taking a closer look at $\mathbb{E}[\phi]$.

Recalling that $\mathbb{V}[\phi] = \mathbb{E}[(\phi - \mathbb{E}[\phi])^2]$, the L2 triangle inequality shows that $\text{dist}(\mathbb{E}[\phi], A)^2 = O(\epsilon)$, and so $\mathbb{E}[\text{dist}(\phi - \mathbb{E}[\phi], A - A)^2] = O(\epsilon)$ by another application of the L2 triangle inequality (recall that $A - A = \{a - b : a, b \in A\}$). This puts us in the same situation as in Lemma 5.5. Applying the lemma to $\phi - \mathbb{E}[\phi]$ with $s = 1$ and $t = 2$, we deduce that

$$\mathbb{V}[\phi] = \mathbb{E}[(\phi - \mathbb{E}[\phi])^2] \leq \epsilon + O(\epsilon^2).$$

For small enough ϵ, the right-hand side is at most 2ϵ.

So far we have managed to show the following:

1. For each $i = 1, \ldots, m$, $\text{dist}(\hat{f}(\{i\}), \Delta)^2 = O(\epsilon)$.
2. $\sum_{i=m+1}^{n} \hat{f}(\{i\}, \Delta)^2 = O(\epsilon)$.

Since $m = O(1)$, this shows that

$$\sum_{i=1}^{n} \text{dist}(\hat{f}(\{i\}), \Delta)^2 = O(\epsilon).$$

To complete the proof, we need to address $\hat{f}(\emptyset)$.

Consider the function h defined by

$$h = \sum_{i=1}^{n} \text{round}(\hat{f}(\{i\}), \Delta) x_i,$$

which satisfies

$$\|h - (f - \hat{f}(\emptyset))\|^2 = \sum_{i=1}^{n} \text{dist}(\hat{f}(\{i\}), \Delta)^2 = O(\epsilon).$$

The function h depends on $m = O(1)$ coordinates, and hence attains values in some finite set E. Therefore

$$\text{dist}(\hat{f}(\emptyset), A - E)^2 \leq \mathbb{E}[(\hat{f}(\emptyset) - (f - h))^2] = O(\epsilon).$$

Consequently, if we set

$$g = h + \text{round}(\hat{f}(\emptyset), A - E)$$

then $\|f - g\|^2 = O(\epsilon)$ by the L2 triangle inequality.

By construction, g is F-valued, where $F = (A - E) + E$. We claim that g is in fact A-valued. Indeed, the L2 triangle inequality shows that

$$\mathbb{E}[\text{dist}(g, A)]^2 = O(\mathbb{E}[\|g - f\|^2] + \mathbb{E}[\text{dist}(f, A)^2]) = O(\epsilon).$$

If g were not A-valued, then since g depends on $O(1)$ variables, $\Pr[g \notin A] = \Omega(1)$. Since g is F-valued, this implies that $\mathbb{E}[\text{dist}(g, A)^2] = \Omega(1)$, and we reach a contradiction for small enough ϵ. We conclude that g is A-valued, completing the proof of Theorem 6.1 for $d = 1$.

6.2 Inductive proof of the Kindler–Safra theorem

In this section we complete the inductive proof of Theorem 6.1, by deducing the theorem for given $d \geq 1$ from the case $d - 1$, using the FKN theorem (the case $d = 1$ of Theorem 6.1).

The idea of the proof is to consider a *random restriction*, leaving a $\sqrt{\epsilon}$ fraction of the coordinates alive, whose effect is to make the function $O(\epsilon)$-close to degree 1 (on average). Applying the FKN theorem, we deduce that the degree 1 Fourier coefficients of the restricted functions are, on average, quantized in aggregate, up to an error of $O(\epsilon)$.

These coefficients are themselves functions of degree $d-1$ *in the parameters of the restriction*, and so applying the case $d-1$ of Theorem 6.1, we are able to deduce that all Fourier coefficients containing exactly one live coordinate are, on average, quantized in aggregate, up to an error of $O(\epsilon)$.

Since a given non-empty Fourier coefficient contains exactly one live coordinate with probability $\Theta(\sqrt{\epsilon})$, the foregoing allows us to conclude that these Fourier coefficients are, on average, quantized in aggregate, up to an error of $O(\sqrt{\epsilon})$. We conclude the proof by handling the empty Fourier coefficient, and by applying the bootstrapping lemma to drive the error down to $O(\epsilon)$.

Random restriction We start with the random restriction argument. Let $S \sim \mu_{\sqrt{\epsilon}}(\{0,1\}^n)$ (we think of S as a set), and let $z \in \{\pm 1\}^{\overline{S}}$ be chosen uniformly at random. We denote by $f|_{\overline{S}\leftarrow z}$ the function on $\{\pm 1\}^S$ obtained by substituting $x_i = z_i$ for all $i \in \overline{S}$ in the Fourier expansion of f.

The Fourier expansion of $f|_{\overline{S}\leftarrow z}$ is supported on subsets of S. For each such $T \subseteq S$,

$$\mathop{\mathbb{E}}_z[\widehat{f|_{\overline{S}\leftarrow z}}(T)^2] = \mathop{\mathbb{E}}_z\left[\left(\sum_{R\subseteq \overline{S}} z_R \hat{f}(T\cup R)\right)^2\right] = \sum_{R\subseteq\overline{S}} \hat{f}(T\cup R)^2 + \sum_{R_1\neq R_2\subseteq\overline{S}} \mathop{\mathbb{E}}_z[z_{R_1\triangle R_2}]\hat{f}(T\cup R_1)\hat{f}(T\cup R_2),$$

where $z_R = \prod_{i\in R} z_i$ and $R_1 \triangle R_2$ is the symmetric difference of R_1 and R_2.

If $R_1 \neq R_2$ then $\mathbb{E}_z[z_{R_1\triangle R_2}] = 0$, and so

$$\mathop{\mathbb{E}}_z[\widehat{f|_{\overline{S}\leftarrow z}}(T)^2] = \sum_{R\subseteq S} \hat{f}(T\cup R)^2.$$

This formula allows us to bound the expected norm of the super-linear part of $f|_{\overline{S}\leftarrow z}$:

$$\mathop{\mathbb{E}}_z[\|(f|_{\overline{S}\leftarrow z})^{>1}\|^2] = \sum_{\substack{|T|>1\\ T\subseteq S}} \sum_{R\subseteq\overline{S}} \hat{f}(T\cup R)^2 = \sum_{|T\cap S|>1} \hat{f}(T)^2.$$

For each particular T, the probability that $|T\cap S| > 1$ is at most $\binom{|T|}{2}\epsilon \le d^2\epsilon$, and so

$$\mathop{\mathbb{E}}_{S,z}[\|(f|_{\overline{S}\leftarrow z})^{>1}\|^2] \le d^2\epsilon \sum_T \hat{f}(T)^2 = O(\epsilon\|f\|^2) = O(\epsilon).$$

We have shown that $f|_{\overline{S}\leftarrow z}$ is very close to degree 1. Also, $\mathbb{E}_{S,z}[\mathrm{dist}(f|_{\overline{S}\leftarrow z}, A)^2] = \epsilon$. Combining these two facts using the L2 triangle inequality, we obtain

$$\mathop{\mathbb{E}}_{S,z}[\mathrm{dist}((f|_{\overline{S}\leftarrow z})^{\le 1}, A)^2] = O(\epsilon). \tag{8}$$

This suggests applying the FKN theorem (the case $d=1$ of Theorem 6.1).

Quantization of non-empty coefficients The FKN theorem states that there exists a finite set B_1 such that for every degree 1 function ϕ,

$$\sum_{|T|\le 1} \mathrm{dist}(\hat{\phi}(T), B_1)^2 = O(\mathbb{E}[\mathrm{dist}(\phi^{\le 1}, A)^2]).$$

Applying this to the functions $f|_{\overline{S}\leftarrow z}$, using (8) we obtain

$$\mathop{\mathbb{E}}_{S,z}\left[\sum_{\substack{|T|\leq 1\\ T\subseteq S}} \operatorname{dist}(\widehat{f|_{\overline{S}\leftarrow z}}(T), B_1)^2\right] = O(\epsilon).$$

Recall that $\widehat{f|_{\overline{S}\leftarrow z}}(T)$ is given by the formula

$$\widehat{f|_{\overline{S}\leftarrow z}}(T) = \sum_{R\subseteq \overline{S}} \hat{f}(T\cup R) z_R,$$

which we think of as a degree $d-|T|$ function in z. For obvious reasons, we will concentrate on the case $|T|=1$. Applying the case $d-1$ of Theorem 6.1 to the functions $\widehat{f|_{\overline{S}\leftarrow z}}(T)$ (as functions on z), with B_1 playing the role of A, we deduce that for some finite set B_2,

$$\mathop{\mathbb{E}}_{S}\left[\sum_{\substack{|T|=1\\ T\subseteq S}} \sum_{R\subseteq \overline{S}} \operatorname{dist}(\hat{f}(T\cup R), B_2)^2\right] = O(\epsilon),$$

which simplifies to

$$\mathop{\mathbb{E}}_{S}\left[\sum_{|T\cap S|=1} \operatorname{dist}(\hat{f}(T), B_2)^2\right] = O(\epsilon),$$

and further to

$$\sum_{T\neq\emptyset} \Pr[|T\cap S|=1] \operatorname{dist}(\hat{f}(T), B_2)^2 = O(\epsilon).$$

The probability that $|T\cap S|=1$ is $|T|\sqrt{\epsilon}(1-\sqrt{\epsilon})^{|T|-1} = \Theta(\sqrt{\epsilon})$ for $T\neq\emptyset$, assuming $\epsilon\leq 1/2$ (say), which we can assume without loss of generality. We conclude that

$$\sum_{T\neq\emptyset} \operatorname{dist}(\hat{f}(T), B_2)^2 = O(\sqrt{\epsilon}). \tag{9}$$

This kind of statement would only allow us to find an $O(\sqrt{\epsilon})$-approximation for f, but fortunately Lemma 5.5 will come to our rescue.

Quantization of empty coefficient Before we can conclude the proof, we need to analyze $\hat{f}(\emptyset)$. To start with, recall that $\|f\|^2 = O(1)$, and so in (9), all but $O(1)$ of the non-empty coefficients are closest to zero. Let J be the set of $O(1)$ coordinates participating in non-empty coefficients which are *not* closest to zero.

We now concentrate on the restrictions $f|_{J\leftarrow w}$. Suppose that $w\in\{\pm 1\}^J$ is chosen at random. As before,

$$\mathop{\mathbb{E}}_{w}\left[\mathbb{V}[f|_{J\leftarrow w}]\right] = \sum_{|T\setminus J|>0} \hat{f}(T)^2 = O(\sqrt{\epsilon}),$$

by the choice of J. Moreover,

$$\mathop{\mathbb{E}}_{w}\left[\mathbb{E}[\operatorname{dist}(f|_{J\leftarrow w}, A)^2]\right] = \epsilon.$$

Markov's inequality shows that the following two conditions hold with probability at least $2/3$ for uniformly random w:

$$\mathbb{V}[f|_{J\leftarrow w}] = O(\sqrt{\epsilon}), \quad \mathbb{E}[\operatorname{dist}(f|_{J\leftarrow w}, A)^2] \leq 3\epsilon.$$

Hence there is some assignment of w for which both occur simultaneously. Recalling that $\mathbb{V}[f|_{J\leftarrow w}] = \mathbb{E}[(f|_{J\leftarrow w} - \mathbb{E}[f|_{J\leftarrow w}])^2]$, the L2 triangle inequality shows that

$$\operatorname{dist}(\mathbb{E}[f|_{J\leftarrow w}], A)^2 = O(\sqrt{\epsilon}).$$

On the other hand,

$$\mathbb{E}[f|_{J\leftarrow w}] = \hat{f}(\emptyset) + \sum_{\emptyset\neq K\subseteq J} w_K \hat{f}(K).$$

The Fourier coefficients on the right-hand side satisfy $\sum_{\emptyset\neq K\subseteq J} \operatorname{dist}(\hat{f}(K), B_2)^2 = O(\sqrt{\epsilon})$, and so the L2 triangle inequality shows that $\operatorname{dist}(\hat{f}(\emptyset), C)^2 = O(\sqrt{\epsilon})$, where C is the following finite set:

$$C = \{a + w_1 b_1 + \cdots + w_{N_J} b_{N_J} : a \in A, w_i \in \{\pm 1\}, b_i \in B_2\},$$

where N_J is a bound on the size of $|J|$.

Culmination of the proof We can now finally describe the approximating function:

$$g = \operatorname{round}(\hat{f}(\emptyset), C) + \sum_{1\leq |T|\leq d} \operatorname{round}(\hat{f}(T), B_2) x_T.$$

The foregoing shows that

$$\|f - g\|^2 = O(\sqrt{\epsilon}).$$

To conclude the proof, we wish to apply Lemma 5.5. To this end, we need to show that the function $f - g$ is ϵ-close to some finite set. Indeed, since g is a junta, the number of values that it can assume belongs to some finite set D. Therefore

$$\mathbb{E}[\operatorname{dist}(f - g, A - D)^2] \leq \mathbb{E}[\operatorname{dist}(f, A)^2] = \epsilon.$$

Applying Lemma 5.5 to $f - g$ (with $s = 2$ and $t = 1$) shows that $\|f - g\|^2 = O(\epsilon)$. By construction, g is E-valued for some finite set E, and so we conclude that g must be A-valued, as in Section 6.1. This concludes the proof of Theorem 6.1 for the given value of d.

7 Junta agreement theorem

Our arguments crucially rely on two agreement theorems, Theorem 3.9 for the p-biased case and Theorem 3.15 for the slice. Both of these theorems immediately follow from more general results proved in the companion paper [DFH17], which do not require the functions ϕ_S to be sparse. Nevertheless, we chose to include the proofs of Theorem 3.9 and Theorem 3.15 for several reasons. First, to make the paper self-contained. Second, the proofs are much simpler than the ones given in [DFH17]. Third, the dependence on d is slightly better.

We prove Theorem 3.15 in Section 7.1, and deduce Theorem 3.9 in Section 7.2.

7.1 On the slice

We deduce Theorem 3.15 from a homogeneous version in which the domain of the functions ϕ_S is $\binom{S}{d}$ rather than $\binom{S}{\leq d}$.

We recall the definition of the distributions ν_ℓ and $\nu_{\ell,t}$. The former is the uniform distribution over the slice $\binom{[n]}{\ell}$. The latter is the distribution of triplets (S_1, S_2, T) sampled as follows: T is a random subset of $[n]$ of size t, and S_1, S_2 are two independent random subsets of $[n]$ of size ℓ containing T.

Theorem 7.1. *Let d, t, ℓ, n, N be positive integers satisfying $d \le t < \ell \le n$, let Σ be an arbitrary set, and let $0 \in \Sigma$ be a distinguished element of Σ. Define $\gamma := \min(t/\ell, 1 - t/\ell)$.*

Suppose that for each $S \in \binom{[n]}{\ell}$ we are given a mapping $\phi_S \colon \binom{S}{d} \to \Sigma$ such that $\phi_S(A) \neq 0$ for at most N many inputs. Let

$$\delta := \Pr_{(S_1,S_2,T)\sim\nu_{\ell,t}}[\phi_{S_1}|_T \neq \phi_{S_2}|_T].$$

For every set $A \in \binom{[n]}{d}$, let $\psi(A)$ be a most common value of $\phi_S(A)$ among all S containing A. Then

$$\Pr_{S\sim\nu_\ell}[\psi|_S \neq \phi_S] = O_{d,\gamma}(N\delta).$$

Furthermore, for every constant $C \ge 1$, the same result holds if $\psi(A)$ satisfies the weaker condition

$$\Pr_{S\sim\nu_\ell}[\phi_S(A) \neq \psi(A) \mid S \supseteq A] \le C \min_{\sigma\in\Sigma} \Pr_{S\sim\nu_\ell}[\phi_S(A) \neq \sigma \mid S \supseteq A].$$

We say that $\psi(A)$ is a C-common value of $\phi_S(A)$. When $C = 1$, this is just a most common value of $\phi_S(A)$.

The exact dependence on d and γ appears below in (12) and the lines following it.

The somewhat mysterious "furthermore" part is required to deduce the theorem on the biased cube.

Proof. We define the following distributions:

- $\nu_{s,A}$ is the uniform distribution over all subsets in $\binom{[n]}{s}$ containing A.
- $\nu_{s,r,A}$ is the distribution of triples (S_1, S_2, T) sampled as follows: $T \sim \nu_{r,A}$, and $S_1, S_2 \sim \nu_{s,T}$ independently.

Our starting point is the following simple observation:

$$\Pr_{S\sim\nu_\ell}[\psi|_S \neq \phi_S] \le \binom{\ell}{d} \Pr_{\substack{S\sim\nu_\ell \\ A\sim\binom{S}{d}}}[\psi(A) \neq \phi_S(A)] = \binom{\ell}{d} \Pr_{\substack{A\sim\nu_d \\ S\sim\nu_{\ell,A}}}[\psi(A) \neq \phi_S(A)].$$

The inequality follows since if $\psi|_S \neq \phi_S$, then at least one of the $\binom{\ell}{d}$ many A's must satisfy $\psi(A) \neq \phi_S(A)$, and so the probability that a random A satisfies $\psi(A) \neq \phi_S(A)$ is at least $1/\binom{\ell}{d}$.

For each particular A, by definition of $\psi(A)$ we have that for each $\sigma \in \Sigma$,

$$\Pr_{S\sim\nu_{\ell,A}}[\psi(A) \neq \phi_S(A)] \le C \Pr_{S\sim\nu_{\ell,A}}[\sigma \neq \phi_S(A)].$$

In particular, choosing $\sigma = \phi_{S_1}(A)$ for $S_1 \sim \nu_{\ell,A}$ and renaming S to S_2, we obtain

$$\Pr_{S\sim\nu_\ell}[\psi|_S \neq \phi_S] \le C\binom{\ell}{d} \Pr_{\substack{A\sim\nu_d \\ S_1,S_2\sim\nu_{\ell,A}}}[\phi_{S_1}(A) \neq \phi_{S_2}(A)].$$

Using the notation

$$\delta_{r,A} = \Pr_{A\sim\nu_d S_1,S_2\sim\nu_{\ell,r,A}}[\phi_{S_1}(A) \neq \phi_{S_2}(A)],$$

this shows that

$$\Pr_{S\sim\nu_\ell}[\psi|_S \neq \phi_S] \leq C\binom{\ell}{d}\mathop{\mathbb{E}}_{A\sim\nu_d}[\delta_{d,A}]. \tag{10}$$

Replace the coupling with an expansion argument?

The goal now is to relate $\delta_{d,A}$ to $\delta_{t,A}$, which we do using a coupling:

- $R \sim \nu_{r,A}$.
- $T_1, T_2 \sim \nu_{s,R}$.
- $S_1 \sim \nu_{\ell,T_1}$; $S_2 \sim \nu_{\ell,T_2}$; $S \sim \nu_{\ell,T_1\cup T_2}$.

The coupling is well-defined if $r \leq s \leq \ell$ and $2s - r \leq \ell$. Fixing r, the largest possible s we can achieve is $\lfloor (r+\ell)/2 \rfloor$.

Observe that $(S, S_1, T_1), (S, S_2, T_2) \sim \nu_{\ell,s,A}$ while $(S_1, S_2, R) \sim \nu_{\ell,r,A}$. Since $\phi_{S_1}(A) \neq \phi_{S_2}(A)$ implies that either $\phi_{S_1}(A) \neq \phi_S(A)$ or $\phi_{S_2}(A) \neq \phi_S(A)$, this shows that

$$\delta_{r,A} \leq 2\delta_{s,A}.$$

How many times do we need to apply this operation to get from $\delta_{d,A}$ to $\delta_{t,A}$? We will keep track of the parameter $\bar{r} = \ell - r$, where r starts at d. Each application of the inequality shrinks $\bar{r}$ by roughly 2: $\bar{r} \mapsto \lceil \frac{\bar{r}}{2} \rceil \leq \frac{\bar{r}+1}{2}$. After k applications, we get to less than $2^{-k}\bar{r} + 1$. Hence after $\lceil \log_2 \frac{\ell-d}{\ell-t} \rceil$ iterations we reach $r = t$. Therefore (10) implies that

$$\Pr_{S\sim\nu_\ell}[\psi|_S \neq \phi_S] \leq 2C\binom{\ell}{d}\frac{\ell-d}{\ell-t}\Pr_{\substack{A\sim\nu_d\\(S_1,S_2,T)\sim\nu_{\ell,t,A}}}[\phi_{S_1}(A) \neq \phi_{S_2}(A)]. \tag{11}$$

Reversing the order of sampling, we see that

$$\Pr_{\substack{A\sim\nu_d\\S_1,S_2,T\sim\nu_{\ell,t,A}}}[\phi_{S_1}(A) \neq \phi_{S_2}(A)] =$$
$$\Pr_{(S_1,S_2,T)\sim\nu_{\ell,t}}[\phi_{S_1}|_T \neq \phi_{S_2}|_T] \cdot \Pr_{\substack{(S_1,S_2,T)\sim\nu_{\ell,t}\\A\in\binom{T}{d}}}[\phi_{S_1}(A) \neq \phi_{S_2}(A) \mid \phi_{S_1}|_T \neq \phi_{S_2}|_T].$$

The first probability is δ by definition. The second probability is at most $2N/\binom{t}{d}$, since at most N entries of each of ϕ_{S_1}, ϕ_{S_2} are non-zero. Plugging this in (11), we conclude

$$\Pr_{S\sim\nu_\ell}[\psi|_S \neq \phi_S] \leq 4C\frac{\binom{\ell}{d}}{\binom{t}{d}}\frac{\ell-d}{\ell-t}N\delta. \tag{12}$$

Since $t/\ell \geq \gamma$ and $t \geq d$, we can bound

$$\frac{\binom{\ell}{d}}{\binom{t}{d}} \leq \frac{\ell^d}{(t-d+1)^d} \leq \left(\frac{t}{t-d+1}\right)^d \gamma^{-d} \leq (d/\gamma)^d.$$

Similarly, since $t/\ell \leq 1-\gamma$, we can bound

$$\frac{\ell-d}{\ell-t} \leq \frac{\ell}{\ell-t} \leq \gamma^{-1}.$$

□

7.2 On the biased cube

We deduce Theorem 3.9 from a homogeneous version, which follows from Theorem 7.1 by a simple reduction.

Let us recall the definition of the distribution $\mu_{p,q}$. It is the distribution of triples (S_1, S_2, T), where $T \sim \mu_q$, and S_1, S_2 are sampled independently by adding to T each $i \notin T$ with probability $r = \frac{p-q}{1-q}$. (The probability r is chosen so that $S_1, S_2 \sim \mu_p$.)

Theorem 7.2. *Let d, n, N be positive integers satisfying $d < n$, let $0 < q < p < 1$, let Σ be an arbitrary set, and let $0 \in \Sigma$ be a distinguished element of Σ. Define $\gamma := \min(q/p, 1 - q/p)$.*

Suppose that for each $S \subseteq \{1, \ldots, n\}$ we are given a mapping $\phi_S \colon \binom{S}{d} \to \Sigma$ such that $\phi_S(A) \neq 0$ for at most N many inputs. Let

$$\delta := \Pr_{(S_1,S_2,T) \sim \mu_{p,q}} [\phi_{S_1}|_T \neq \phi_{S_2}|_T].$$

For every set $A \in \binom{[n]}{d}$, let $\psi(A)$ be a most common value of $\phi_S(A)$ among all S containing A. Then

$$\Pr_{S \sim \nu_\ell} [\psi|_S \neq \phi_S] = O_{d,\gamma}(N\delta).$$

Proof. Let Z be a large integer. Let $n_Z = Zn$, $\ell_Z = \lfloor Zp \rfloor$, and $t_Z = \lfloor Zq \rfloor$. We assume that Z is large enough so that $\ell_Z > t_Z$. Note that $d \leq t_Z < \ell_Z \leq n_Z$, and $\gamma_Z = \gamma + O(1/Z)$.

Let $\mu_{p,Z}$ be the following distribution on $\{0,1\}^n$: sample $S \sim \nu_{\ell_Z}$, and take $S \cap [n]$. Clearly $\mu_{p,Z} \to \mu_p$ as $Z \to \infty$ (since $\{0,1\}^n$ is finite, this is true with respect to any reasonable metric on distributions). Similarly, if $\mu_{p,q,Z}$ is the distribution formed by sampling $(S, T_1, T_2) \sim \nu_{\ell_Z, t_Z}$ and taking $(S \cap [n], T_1 \cap [n], T_2 \cap [n])$, then $\mu_{p,q,Z} \to \mu_{p,q}$ as $Z \to \infty$.

For each $S \in \binom{[n_Z]}{\ell_Z}$, define a mapping $\phi_{Z,S} \colon \binom{S}{d} \to \Sigma$ by

$$\phi_{Z,S}(A) = \begin{cases} \phi_{S \cap [n]}(A) & \text{if } A \subseteq [n], \\ 0 & \text{otherwise.} \end{cases}$$

By assumption, $\phi_{Z,S}(A) \neq 0$ for at most N many inputs. Define

$$\delta_Z = \Pr_{(S_1,S_2,T) \sim \nu_{\ell_Z, t_Z}} [\phi_{Z,S_1}|_T \neq \phi_{Z,S_2}|_T],$$

and notice that $\phi_{Z,S_1}|_T \neq \phi_{Z,S_2}|_T$ iff $\phi_{S_1 \cap [n]}|_{T \cap [n]} \neq \phi_{S_2 \cap [n]}|_{T \cap [n]}$. Since $\mu_{p,q,Z} \to \mu_{p,q}$, this implies that $\delta_Z \to \delta$.

For a set $A \in \binom{[n]}{d}$, recall that $\psi(A)$ is a most common value of $\phi_S(A)$ among all S containing A. Since $\mu_{p,Z} \to \mu_p$, for every $\sigma \in \Sigma$ we have

$$\Pr_{S \sim \nu_{\ell_Z}} [\phi_{Z,S}(A) \neq \sigma \mid S \supseteq A] \to \Pr_{S \sim \mu_p} [\phi_S(A) \neq \sigma \mid S \supseteq A].$$

In particular, for large enough Z, the ratio between the two sides lies in the range $[1/2, 2]$. Since $\psi(A)$ is a most common value of $\phi_S(A)$, for large enough Z we have

$$\Pr_{S\sim\nu_{\ell_Z}}[\phi_{Z,S}(A) \neq \psi(A) \mid S \supseteq A] \leq 2 \Pr_{S\sim\mu_p}[\phi_S(A) \neq \psi(A) \mid S \supseteq A]$$
$$= 2 \max_{\sigma\in\Sigma} \Pr_{S\sim\mu_p}[\phi_S(A) \neq \sigma \mid S \supseteq A] \leq 4 \max_{\sigma\in\Sigma} \Pr_{S\sim\nu_{\ell_Z}}[\phi_{Z,S}(A) \neq \sigma \mid S \supseteq A].$$

Therefore, if we define $\psi_Z(A)$ to be $\psi(A)$ if $A \subseteq [n]$ and 0 otherwise, then $\psi_Z(A)$ is a 4-common value of $\phi_{Z,S}(A)$. Theorem 7.1 shows that

$$\Pr_{S\sim\nu_{\ell_Z}}[\psi_Z|_S \neq \phi_{Z,S}] = O_{d,\gamma_Z}(N\delta_Z).$$

Since $\psi_Z|_S \neq \phi_{Z,S}$ iff $\psi_{S\cap[n]} = \psi_Z|_{S\cap[n]} \neq \phi_{Z,S}|_{S\cap[n]} = \phi_{S\cap[n]}$, we have

$$\Pr_{S\sim\nu_{\ell_Z}}[\psi|_{S\cap[n]} \neq \phi_{S\cap[n]}] = O_{d,\gamma_Z}(N\delta_Z).$$

The dependence of the big O on γ_Z is continuous. Therefore, since $\gamma_Z \to \gamma$, $\delta_Z \to \delta$, and $\mu_{p,Z} \to \mu_p$, taking a large enough Z, we conclude the statement of the theorem. □

References

[BHP01] Roger C. Baker, Glyn Harman, and János Pintz. *The difference between consecutive primes, II*. Proc. Lond. Math. Soc., 83(3):532—562, 2001. 19

[BK97] Jean Bourgain and Gil Kalai. *Influences of variables and threshold intervals under group symmetries*. Geom. Funct. Anal., 7(3):438–461, 1997. 2

[Bou99] Jean Bourgain. *On sharp threshold of monotone properties*. J. Amer. Math. Soc., 12(4):1017–1054, 1999. (Appendix to *E. Friedgut, Sharp Thresholds of Graph Properties, and the k-SAT Problem*). 2

[Bou02] ———. *On the distribution of the Fourier spectrum of Boolean functions*. Israel J. Math., 131(1):269–276, 2002. 2

[BW02] Harry Buhrman and Ronald de Wolf. *Complexity measures and decision tree complexity: a survey*. Theoret. Comput. Sci., 288(1):21–43, 2002. (Special issue on Complexity and Logic, Preliminary version in International Workshop Collegium Logicum 1998). 25

[CHS20] John Chiarelli, Pooya Hatami, and Michael E. Saks. *An asymptotically tight bound on the number of relevant variables in a bounded degree Boolean function*. Combinatorica, 40(2):237–244, 2020. arXiv:1801.08564. 29

[CST17] Gil Cohen, Amir Shpilka, and Avishay Tal. *On the degree of univariate polynomials over the integers*. Combinatorica, 37(3):419–464, 2017. (Preliminary version in *3rd ITCS*, 2012). eccc:2011/TR11-002. 19

[DFH17] Irit Dinur, Yuval Filmus, and Prahladh Harsha. *Agreement tests on graphs and hypergraphs*, 2017. (manuscript). arXiv:1711.09426, eccc:2017/TR17-181. 5, 38

[DFH19] ———. *Analyzing Boolean functions on the biased hypercube via higher-dimensional agreement tests*. In *Proc. 30th Annual ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 2124–2133. 2019. arXiv:1711.09428,1711.09426, eccc:2017/TR17-180,2017/TR17-181. 5

[Dun76] Charles Dunkl. *A Krawtchouk polynomial addition theorem and wreath products of symmetric groups*. Indiana Univ. Math. J., 25:335–358, 1976. 20, 22

[FI19] Yuval Filmus and Ferdinand Ihringer. *Boolean constant degree functions on the slice are juntas*. Discret. Math., 342(12), 2019. arXiv:1801.06338. 19, 27

[Fil16a] Yuval Filmus. *Friedgut-Kalai-Naor theorem for slices of the Boolean cube*. Chic. J. Theoret. Comput. Sci., 2016(14), 2016. arXiv:1410.7834. 2, 3, 5, 19, 24

[Fil16b] ———. *An orthogonal basis for functions over a slice of the Boolean hypercube*. Electron. J. Combin., 23(1):P1.23, 2016. arXiv:1406.0142. 20, 22

[FK96] Ehud Friedgut and Gil Kalai. *Every monotone graph property has a sharp threshold*. Proc. Amer. Math. Soc., 124(10):2993–3002, 1996. 2

[FKMW18] Yuval Filmus, Guy Kindler, Elchanan Mossel, and Karl Wimmer. *Invariance principle on the slice*. ACM T. Comput. Theory, 10(3):11:1–11:37, 2018. (Preliminary version in *31st Comput. Complexity Conf.*). arXiv:1504.01689. 20

[FKN02] Ehud Friedgut, Gil Kalai, and Assaf Naor. *Boolean functions whose Fourier transform is concentrated on the first two levels and neutral social choice*. Adv. Appl. Math., 29(3):427–437, 2002. 2, 33, 34

[FM19] Yuval Filmus and Elchanan Mossel. *Harmonicity and invariance on slices of the Boolean cube*. Probab. Theory Related Fields, 175:721—782, 2019. (Preliminary version in *31st Comput. Complexity Conf.*). arXiv:1507.02713. 20, 22, 28

[Fri98] Ehud Friedgut. *Boolean functions with low average sensitivity depend on few coordinates*. Combinatorica, 18(1):27–35, 1998. 1, 2

[Fri99] ———. *Sharp thresholds of graph properties, and the k-SAT problem*. J. Amer. Math. Soc., 12(4):1017–1054, 1999. 2

[GR97] Joachim von zur Gathen and James R. Roche. *Polynomials with two values*. Combinatorica, 17(3):345–362, 1997. 5, 19

[Hat12] Hamed Hatami. *A structure theorem for Boolean functions with small total influences*. Ann. of Math., 176(1):509–533, 2012. 1, 2

[Kin03] Guy Kindler. *Property Testing, PCP and Juntas*. Ph.D. thesis, Tel Aviv University, 2003. eccc:2016/TR16-007. 1, 11, 30, 33, 34

[KK20] Nathan Keller and Ohad Klein. *A structure theorem for almost low-degree functions on the slice*. Israel J. Math., 240(1):179–220, 2020. arXiv:1901.08839. 20, 31, 32

[Lif18] Noam Lifshitz. Private communication, 2018. 20

[NS94] Noam Nisan and Mario Szegedy. *On the degree of Boolean functions as real polynomials*. Comput. Complexity, 4:301–313, 1994. (Preliminary version in *24th STOC*, 1992). 19, 29

[O'D14] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. doi:10.1017/CBO9781139814782. 1, 29

[Sta91] Richard P. Stanley. *Some applications of algebra to combinatorics*. Discrete Appl. Math., 34(1):241–277, 1991. 22

[Wel19] Jake Wellens. *A tighter bound on the number of relevant variables in a bounded degree Boolean function*, 2019. (manuscript). arXiv:1903.08214. 29