# Finding Skewed Subcubes Under a Distribution

**Parikshit Gopalan**
VMware Research, Palo Alto, CA, USA
pgopalan@vmware.com

**Roie Levin**[1]
Carnegie Mellon University, Pittsburgh, PA, USA
roiel@cs.cmu.edu

**Udi Wieder**
VMware Research, Palo Alto, CA, USA
uwieder@vmware.com

## Abstract

Say that we are given samples from a distribution $\psi$ over an $n$-dimensional space. We expect or desire $\psi$ to behave like a product distribution (or a $k$-wise independent distribution over its marginals for small $k$). We propose the problem of enumerating/list-decoding all large subcubes where the distribution $\psi$ deviates markedly from what we expect; we refer to such subcubes as skewed subcubes. Skewed subcubes are certificates of dependencies between small subsets of variables in $\psi$. We motivate this problem by showing that it arises naturally in the context of algorithmic fairness and anomaly detection.

In this work we focus on the special but important case where the space is the Boolean hypercube, and the expected marginals are uniform. We show that the obvious definition of skewed subcubes can lead to intractable list sizes, and propose a better definition of a minimal skewed subcube, which are subcubes whose skew cannot be attributed to a larger subcube that contains it. Our main technical contribution is a list-size bound for this definition and an algorithm to efficiently find all such subcubes. Both the bound and the algorithm rely on Fourier-analytic techniques, especially the powerful hypercontractive inequality.

On the lower bounds side, we show that finding skewed subcubes is as hard as the sparse noisy parity problem, and hence our algorithms cannot be improved on substantially without a breakthrough on this problem which is believed to be intractable. Motivated by this, we study alternate models allowing query access to $\psi$ where finding skewed subcubes might be easier.

**2012 ACM Subject Classification** Mathematics of computing → Probabilistic algorithms

**Keywords and phrases** Fourier Analysis, Anomaly Detection, Algorithmic Fairness, Probability, Unsupervised Learning

**Digital Object Identifier** 10.4230/LIPIcs.ITCS.2020.84

**Related Version** A full version of the paper is available at `https://arxiv.org/abs/1911.07378`.

## 1 Introduction

Assume that we observe samples from a distribution $\psi$ over points in $n$-dimensional space $\mathcal{D}^n$. Our prior belief is that each attribute has a marginal distribution $\mu_i$ and that the various attributes are nearly independent (or at least $k$-wise independent for small $k$), hence $\psi$ is close to the product distribution $\mu = \prod_i \mu_i$. Our goal is to find significant deviations between our hypothesis $\mu$ and the observed distribution $\psi$, manifested as significant dependencies between small sets of variables. The distribution $\mu$ might represent either a prior model for

[1] Work done while an intern at VMware Research.


$\psi$, or it might be represent a target distribution that we wish $\psi$ to be close to. This problem arises naturally in several machine learning applications as we detail in Section 1.1, but first we formulate the problem with more detail.

To formulate a precise statement, we first define the notion of subcubes. Assume that $\mathcal{D}$ is ordered and bounded, the two canonical examples are $\mathcal{D}^n = \{0,1\}^n$ and $\mathcal{D}^n = [0,1]^n$. Let $K \subseteq [n]$ be a set of $k$ coordinates. For $x \in \mathcal{D}^n$, $x_K$ denotes the projection of $x$ onto coordinates in $K$. For each $j \in K$, let $I_j \subsetneq \mathcal{D}$ be an interval in $\mathcal{D}$. We call the set of points $C = \{x \in \mathcal{D}^n : x_K \in \prod_{j \in K} I_j\}$ a subcube of codimension $k$. We have

$$\mu(C) := \Pr_{\mathbf{x} \sim \mu}[\mathbf{x} \in C] = \prod_{j \in K} \Pr_{\mathbf{x}_j \sim \mu_j}[\mathbf{x}_j \in I_j] = \prod_{j \in K} \mu_j(I_j).$$

If we similarly define $\psi(C) := \Pr_{\mathbf{x} \sim \psi}[\mathbf{x} \in C]$, then our goal is to find subcubes such that $|\mu(C) - \psi(C)| \geq \gamma$. Motivated by our applications, we add two more desiderata to our problem formulation (that will be justified shortly): we restrict to large subcubes, and we want algorithms that enumerate all subcubes that satisfy our conditions.

One way to restrict to large subcubes is to only consider subcubes with $\mu(C) \geq \eta$ for some $\eta \in [0,1]$. Alternately, we could bound the codimension by $k$. The advantage of the latter is that we only need that $\mu$ is $k$-wise independent for the equality $\mu(C) = \prod_{j \in K} \mu_j(I_j)$ to hold. In the discrete case $\mathcal{D}^n = \{0,1\}^n$, the two notions coincide since $\mu(C) = 2^{-k}$ for subcubes of codimension $k$.

Rather then phrasing this as an optimization question where the goal is to find the subcube that maximizes the deviation $\gamma$, our goal will be to come up with a *list-decoding* style algorithm that enumerates over all subcubes of codimension $k$ such that $|\mu(C) - \psi(C)| \geq \gamma$.

In addition to being a natural algorithmic question in its own right, this problem comes up in recent work in machine learning, on anomaly detection and fairness.

## 1.1 Motivation

### Fairness in Machine Learning

Assume there is a base population $P$ of individuals, each described by $n$ attributes. We naturally view $P$ as inducing a distribution $\mu$ on the attribute space $\mathcal{D}^n$. Suppose that small subsets of the attributes are nearly independent, so that $\mu$ is close to being $k$-wise independent for some $k$ which is small compared to $n$. We are given a distribution $\psi$ over this population. Our goal is to discover significant biases in the distribution that are not present in the original population $P$. For instance the population $P$ might be the set of students that apply to a university, and $\psi$ might represent the set of successful applicants. Or $P$ might be the training data for a machine learning algorithm while $\psi$ represents the misclassified inputs. The latter setting has received a fair amount of attention in the context of algorithmic bias and fairness in Machine learning, where the most commonly studied notion is that of intersectionality bias [4]: we are interested in biases where we restrict the values of some small subset of attributes, which are typically discrete. See for instance a recent study showing that facial recognition software has higher error rates for women of color [3]. Our motivation for considering subcubes is that it captures intersectionality in the discrete setting.

Enumerating over all subcubes is more appropriate than optimization in this setting since not all intersectionalities might be equally important. The fact that college applications submitted during certain days of the week are less likely to be accepted might not be as significant as the fact that certain zipcodes are less likely to be accepted; even if the deviation

is lower in the latter case. We ask for algorithms that enumerate over all biased subcubes and leave it to subject experts to decide how interesting these are, just as in list-decoding we do not worry about how the receiver chooses from the list of possible codewords returned by the decoder. Another reason to favor enumeration is that in real-world datasets, we may not expect $\psi$ to be truly $k$-wise independent; we might expect correlations between certain sets of attributes. But even so, an exhaustive list of significant correlations might lead us to discover interesting new properties of the distribution and refine our model for $\psi$. The restriction to subcubes of bounded codimension is natural since intersectionalities of few attributes are more interesting.

## Anomaly Detection

Anomaly detection is a ubiquitous unsupervised learning problem [5]. Isolation based methods for anomaly detection have proven to be extremely effective in practice [15, 7, 11]. Building on this, the recent work of [10] proposes an approach to anomaly detection based on a notion called Partial Identification. It assigns a score denoted $\mathrm{PIDScore}(x, P)$ to each point $x \in P$ which measures how easy it is to distinguish $x$ from other points in $P$. They give a heuristic to compute $\mathrm{PIDScore}(x, P)$, and show that the resulting anomaly detection algorithm outperforms several popular anomaly detection methods, across a broad range of benchmarks.

Formally, given a set of points $P \subseteq \mathcal{D}^n$ and a subcube $C \in \mathcal{D}^n$, define the sparsity of $C$ as

$$\rho(C) = \frac{\mathrm{vol}(C)}{|C \cap P|}$$

The PIDScore of a point $x \in P$ is the maximum value of $\rho(C)$ over all subcubes that contain it.

$$\mathrm{PIDScore}(x, P) = \max_{C \ni x} \rho(C).$$

Anomalous points are those for which $\mathrm{PIDScore}(x, P) \geq t$ for some threshold $t$. Equivalently, it suffices to find all $C$ such that $\rho(C) \geq t$, and then take all the points contained in them.

To relate this to our problem, let us take $\mu$ to be the uniform measure over $\mathcal{D}^n$ and $\psi$ to be the measure induced by $P$. Rescaling $\rho$ by a factor of $|P|/\mathrm{vol}(\mathcal{D}^n)$, we get

$$\rho'(C) = \frac{\mathrm{vol}(C)}{\mathrm{vol}(\mathcal{D}^n)} \frac{|P|}{|C \cap P|} \approx \frac{\mu(C)}{\psi(C)}.^{2}$$

If we also scale the threshold $t$ by the same factor, then the set of outliers stays the same. But $\rho'(C) \geq t'$ implies $\psi(C) \leq \mu(C)/t'$, hence

$$\frac{\mu(C) - \psi(C)}{\mu(C)} \geq 1 - \frac{1}{t'}.$$

Thus this is an instance of the problem that we consider, where our goal is to find non-empty subcubes that are underrepresented in $\psi$, when compared to $\mu$. Enumeration over all sparse subcubes is natural in this setting, since we wish to list all points with high scores.

[2] Actually $|C \cap P|/|P| = \psi(C)$ only in expectation, but since subcubes have small VC dimension, we get tight concentration.

## 2 Our Results

In this paper, we focus on the case when $\mathcal{D}^n = \{\pm 1\}^n$ and $\mu$ is the uniform distribution. We believe that several of our techniques apply to more general product distributions. A subcube of codimension $k$ is obtained by restricting the values of some subset $K$ of coordinates. For subcubes $C \subseteq D$ we refer to $C$ as a child of $D$ and $D$ as a parent of $C$.

As a warm-up we first consider the following problem:

▶ **Problem 1** (Finding skewed subcubes). *Given sample access to a distribution $\psi$ over $\{\pm 1\}^n$ and $\gamma \in (0, 2^k - 1]$ find all subcubes $C$ with codimension $j \leq k$ such that*

$$\left| \frac{\Pr_{\mathbf{x}\sim\psi}[\mathbf{x} \in C] - 2^{-j}}{2^{-j}} \right| \geq \gamma.$$

There is a trivial $\tilde{O}(n^k)$ algorithm that enumerates over all subcubes. To beat this naive bound, we first need to bound the list-size of the output, or rather a bound on the number of skewed subcubes. However, we show in Lemma 12 that there exist distributions where the number of skewed subcubes is $\Omega((n/k)^k)$, which is not far from the trivial upper bound.

The proof of Lemma 12 demonstrates that one source for the abundance of skewed subcubes is that skew is easily inherited by children from their parents: if a subcube $C$ of codimension $j$ is skewed, for every choice of $k - j$ additional coordinates, by simple averaging, there is at least one restriction that results in a skewed subcube. So even if we consider the uniform distribution over points with $x_1 = 1$, there are $\Omega(n^{k-1})$ skewed subcubes by this definition, while really the only interesting subcube is the $x_1 = 1$ subcube. Our first contribution is a definition which captures only those subcubes that do not inherit their skew from a parent.

▶ **Problem 2** (Finding minimal skewed subcubes). *Given sample access to a distribution $\psi$ over $\{\pm 1\}^n$, $\gamma \in (0, 2^k - 1]$ and $\epsilon \in (0, 1)$ find all subcubes $C$ with codimension $j \leq k$ such that*

$$\left| \frac{\Pr_{\mathbf{x}\sim\psi}[\mathbf{x} \in C] - 2^{-j}}{2^{-j}} \right| \geq \gamma.$$

*and for every parent $C' \supsetneq C$ of codimension $i$,*[3]

$$\left| \frac{\Pr_{\mathbf{x}\sim\psi}[\mathbf{x} \in C] - 2^{-i}}{2^{-i}} \right| \leq \gamma(1 - \epsilon).$$

We refer to such a subcube as a $(\gamma, \epsilon)$-minimal skewed subcube. This notion is motivated by our applications: if we already know that $\Pr_{\mathbf{x}\sim\psi}[(\mathbf{x}_1 = 1) \wedge (\mathbf{x}_2 = 1)] = 3/4$ (rather than $1/4$), then knowing that $\Pr_{\mathbf{x}\sim\psi}[(\mathbf{x}_1 = 1) \wedge (\mathbf{x}_2 = 1) \wedge (\mathbf{x}_3 = 1)] = 3/8$ should not surprise us, given our prior.

A natural question to ask is whether focusing in minimal skewed subcubes suffices to make the problem (or at least the list size) more tractable. Our second contribution is a bound on the number of minimal skewed subcubes which is independent of the dimension $n$. Instead we have a dependence on the max norm of the probability distribution defined below.

[3] The formal definition of minimal skewed subcubes (Definition 13) is a little more involved, we only care about those parents of $C$ which are skewed the same way as $C$.

Given a distribution $\psi$ on $\{\pm 1\}^n$, let $\|\psi\|_\infty := 2^n \cdot \max_{x\in\{\pm1\}^n} \Pr_{\mathbf{x}\sim\psi}[\mathbf{x} = x]$. The parameter $\|\psi\|_\infty$ lies in the range $[1, 2^n]$ and is a measure of how well-spread the distribution is. It is referred to as the smoothness of a distribution in the literature on boosting, and is closely related to min-entropy. The uniform distribution has $t = 1$, whereas $t = 2^n$ when the entire distribution is concentrated on a single point.

▶ **Theorem 1.** *For a distribution $\psi$ on $\{\pm1\}^n$, the number of $(\gamma, \epsilon)$-minimal skewed subcubes of codimension at most $k$ is bounded by $k^{O(k)} \left(\ln(e\,\|\psi\|_\infty)\,\mathrm{poly}(1/\epsilon, 1/\gamma)\right)^k$.*

For constant $\epsilon, \gamma$, the asymptotic dependence on $n$ is never worse than $O(n^k)$, which happens when $\psi$ is concentrated on a point. But when $\|\psi\|_\infty = O(1)$, the above bound is $O_k(1)$ and when $\|\psi\|_\infty = \mathrm{poly}(n)$, the bound is $O_k(\ln(n)^k)$ improving substantially over the $O(n^k)$ bound.

There are two key elements in the proof of Theorem 1. We first use a novel Fourier based algorithm to reduce the problem to that of finding large, low-degree Fourier coefficients in a series of restrictions of the distribution $\psi$ to various subcubes. We then use the powerful hypercontractive inequality to bound the number of such coefficients in any distribution in terms of $\|\psi\|_\infty$. This latter bound generalizes the level-$k$ inequalities for indicators of small sets in the Boolean hypercube [16, Chapter 9], and the proof follows similar lines. We also construct distributions showing that for various values of $\epsilon, \gamma$, the dependency of $(\ln(\|\psi\|_\infty))^k$ is optimal. The distributions are constructed using the Tribes function and BCH codes.

We now turn to the algorithmic problem of finding the list of minimal skewed subcubes. We observe that even when the list-size is constant, there is a significant algorithmic barrier to a $n^{o(k)}$ algorithm, namely the $k$-sparse noisy parity problem [8, 18]. In this problem, we are given points $x$ and labels $y$ which are the XOR of some $k$-subset $S$ with random noise of rate $\eta$ added. There is a simple reduction from this problem to finding skewed subcubes, if we consider the distribution of $(x, y) \in \{\pm1\}^{n+1}$ the only skewed subcubes involve the coordinates $S \cup \{n + 1\}$.

▶ **Theorem 2.** *For $\eta \in (0, 1/2)$, an algorithm that given a distribution $\psi$ and $k$ can find a $(1 - 2\eta, 1)$-minimal skewed subcube of co-dimension $k$ in time $T(n, k, \eta)$ can be used to solve the $k$-sparse noisy parity problem with noise rate $\eta$ in time $T(n, k, \eta)$.*

Given this reduction, there are two *lower bounds* on the running time of any list-decoder: the list-size given in Theorem 1, and the running time of the best known algorithm for the $k$-sparse noisy parity problem, which is $O(n^{0.8k})$ due to [18]. We give an algorithm that nearly gets the sum of these two bounds.

▶ **Theorem 3.** *For any measure $\psi$ on $\{\pm1\}^n$, integer $k \le n$, and parameters $0 \le \gamma \le 2^k - 1$ and $0 \le \epsilon \le 1$, there are algorithms that return all $(\gamma, \epsilon)$-minimal skewed subcubes of codimension at most $k$ in time*

$$\tilde{O}\left(n^{0.8k}\right) + \tilde{O}\left(n^{k/3}\right) \cdot \frac{k^{O(k)}}{(\epsilon\gamma)^{4/\lambda+2k}} \left(\ln\left(\frac{e\,\|\psi\|_\infty}{\epsilon\gamma}\right)\right)^k$$

Finally, to circumvent the noisy parity problem, we consider stronger models where we have query access to the distribution $\psi$: in addition to random samples, we can also query the value of $\psi(x)$ for any $x \in \{\pm1\}^n$. The noisy parity problem becomes trivial to solve once one has query access. In this model, we are able to get an algorithm whose running time is is $\mathrm{poly}(n, \|\psi\|_\infty)$. Thus when $\|\psi\|_\infty < n^{\alpha k}$ for some $\alpha > 0$, this improves over the trivial algorithm. We show some dependence on $\|\psi\|_\infty$, possibly of the form $\ln(\|\psi\|_\infty)^k$ is inherent even in the query model, by constructing a distribution $\psi$ (with large $\|\psi\|_\infty$) where the query model and random samples model are equivalent, and where finding skewed subcubes lets us solve the $k$-sparse noisy parity problem.

## 2.1 Related Work

In nearby work, [1] study the problem of testing whether a distribution is $\delta$ close in statistical distance to $(\epsilon, k)$-wise uniform. In our language, a distribution $D$ is said to be $(\epsilon, k)$-wise uniform if all subcubes of codimension $j \leq k$ have skew no greater than $2^j\epsilon$. They provide a sample complexity upper bound of $O((k \log n)/\epsilon^2\delta^2)$. They then provide evidence, based on the conjectured hardness of finding planted cliques, that no polynomial in $n$ time algorithm for this problem exists (one can also base this hardness on sparse noisy parity, as we do here). Indeed, their testing algorithm essentially reduces the problem to the optimization version: find the subcube of codimension $k$ such that the skew is maximized.

Fourier analytic techniques have found widespread use in a variety of supervised learning problems under the uniform distribution [16]. Our work differs from this in that the problem we consider is an unsupervised learning problem, and that we use Fourier analysis over the uniform distribution to reason about the deviation from an arbitrary distribution. In this aspect, our work is similar to the work of [1, 17].

Finally, there have been a line of recent results in machine learning which have a list-decoding flavor to them, see for instance [6, 12].

**Outline of the paper.** Section 3 introduces definitions and notation. Section 4 contains Fourier analytic results that are required for our results. Section 5 proves our main combinatorial bounds on the number of skewed subcubes, and gives examples that show these bounds are tight. Section 6 describes the efficient algorithm for enumerating minimal skewed subcubes. Section 7 gives lower bounds due to the reduction from the noisy parity problem. Section 8 considers the problem in the membership query model. Some proofs are deferred from the main body to the Appendix 9.

# 3 Definitions

In this section we present basic definitions and facts.

### Distributions

We denote the $n$-dimensional Hamming cube by $\{\pm 1\}^n$. Given a probability distribution $\psi$ on $\{\pm 1\}^n$, it is convenient to identify it with the probability measure $\psi : \{\pm 1\}^n \to \mathbb{R}^{\geq 0}$ which satisfies

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim\{\pm 1\}^n}[\psi(\mathbf{x})] = 1.$$

We write $\mathbf{x} \sim \psi$ to denote $\mathbf{x}$ is a random variable with the distribution

$$\Pr_{\mathbf{x}\sim\psi}[\mathbf{x} = x] = \frac{\psi(x)}{2^n}$$

Henceforth, we will interchangeably refer to $\psi$ as a distribution and a measure. We will use $\mu$ to denote the uniform distribution over $\{\pm 1\}^n$, where $\mu(x) = 1$ for all $x \in \{\pm 1\}^n$. Given functions $f, g : \{\pm 1\}^n \to \mathbb{R}$, we define their inner product by $\langle f, g\rangle := \mathbb{E}_{\mathbf{x}\sim\mu}[f(\mathbf{x})g(\mathbf{x})]$. We define $\|f\|_p := \mathbb{E}_{\mathbf{x}\sim\mu}[f(\mathbf{x})^p]^{1/p}$. For two probability measures $\psi, \theta$, we have

$$\langle \psi, \theta\rangle = \sum_{x\in\{\pm 1\}^n} \frac{\psi(x)\theta(x)}{2^n} = \mathop{\mathbb{E}}_{\mathbf{x}\sim\psi}[\theta(\mathbf{x})] = \mathop{\mathbb{E}}_{\mathbf{x}\sim\theta}[\psi(\mathbf{x})].$$

Let $A \subseteq \{\pm 1\}^n$ and let $\alpha = |A|/2^n$ denote its fractional density. We use $\mu_A$ to denote the uniform distribution over $A$. The corresponding measure is defined as

$$\mu_A(\mathbf{x}) = \begin{cases} \frac{1}{\alpha} & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

For a distribution $\psi$, we define

$$\|\psi\|_\infty = \max_{x \in \{\pm 1\}^n} \psi(x)$$

It follows from the definition that

$$\|\psi\|_\infty = 2^n \max_{x \in \{\pm 1\}^n} \Pr_{\mathbf{x} \sim \psi}[\mathbf{x} = x] = \max_{x \in \{\pm 1\}^n} \frac{\Pr_{\mathbf{x} \sim \psi}[\mathbf{x} = x]}{\Pr_{\mathbf{x} \sim \mu}[\mathbf{x} = x]}$$

A bound on $\|\psi\|_\infty$ implies that no point is too likely.

### Subcubes

A subcube is a subset of $\{\pm 1\}^n$ obtained by fixing some subset of bits to a particular value. Formally, a subcube $C \subseteq \{\pm 1\}^n$ is specified by a pair $(K, y)$ where $K \subseteq [n]$ and $y \in \{\pm 1\}^K$. We have

$$C = \{x \in \{\pm 1\}^n \text{ s.t. } x_i = y_i \ \forall\ i \in K\}.$$

We refer to coordinates in $K$ as the fixed coordinates of $C$, and to the rest as the free coordinates of $C$. For $C = (K, y)$ we define the codimension of $C$ to be $|K|$ and denote it $\mathrm{codim}(C)$. We use $\mathcal{C}^{\leq k}$ to denote the set of all subcubes of codimension at most $k$. By Equation (1), $\mu_C$ the uniform measure over $C$ is given by

$$\mu_C(x) = \begin{cases} 2^{\mathrm{codim}(C)} & \text{if } x \in C \\ 0 & \text{if } x \notin C \end{cases}$$

For subcubes $C = (K, y), D = (L, z)$, we have $D \subset C$ iff $K \subset L$ and $z_i = y_i$ for all $i \in K$. We refer to $D$ as a child of $C$ and $C$ as a parent of $D$.

▶ **Definition 4** (Restriction). *For a distribution $\psi$ on $\{\pm 1\}^n$ and a subcube $C \subseteq \{\pm 1\}^n$ such that $\psi$ assigns non-zero probability to $C$ we define $\psi|_C : C \to \mathbb{R}^{\geq 0}$, the restriction of $\psi$ to $C$, as*

$$\psi|_C(x) = \frac{\psi(x)}{\langle \psi, \mu_C \rangle}.$$

Since $\langle \psi, \mu_C \rangle = \mathbb{E}_{x \sim \mu_C}[\psi(x)]$, $\psi$ assings non-zero probaility to $C$ iff $\langle \psi, \mu_C \rangle > 0$. The restriction is itself a legal probability measure; it satisfies $\mathbb{E}_{x \in C}[\psi|_C] = 1$. This definition immediately implies the relationship:

▶ **Fact 5.** *For a distribution $\psi$ on $\{\pm 1\}^n$ and a subcube $C \subset \{\pm 1\}^n$*

$$\|\psi|_C\|_\infty = \frac{\|\psi\|_\infty}{\langle \psi, \mu_C \rangle}$$

▶ **Lemma 6.** *Given subcubes $C$ and $D$ such that $D \subseteq C \subseteq \{\pm 1\}^n$, and a density function $\psi$, it holds that:*

$$\langle \psi, \mu_D \rangle = \langle \psi, \mu_C \rangle \cdot \langle \psi|_C, \mu_D|_C \rangle$$

## 3.1 Skewed subcubes

▶ **Definition 7** (Skew). *We define the **skew** of a subcube $C$ with respect to measure $\psi$ as*

$$\textsc{Skew}_\psi(C) = \langle \psi, \mu_C \rangle - 1$$

The next two lemmas state some simple facts about the skew of subcubes. First we show the skew of a subcube measures the deviation of the measure on the subcube from the uniform distribution.

▶ **Lemma 8.** *Let $\mathrm{codim}(C) = k$. We have*

$$\begin{aligned}\textsc{Skew}_\psi(C) &= 2^k \Pr_{\mathbf{x}\sim\psi}[\mathbf{x} \in C] - 1 \\ &= \frac{1}{\Pr_{x\sim\mu}[\mathbf{x} \in C]} \left( \Pr_{x\sim\psi}[\mathbf{x} \in C] - \Pr_{x\sim\mu}[\mathbf{x} \in C] \right).\end{aligned}$$

▶ **Corollary 9.** *For any distribution $\psi$, $\textsc{Skew}_\psi(C)$ lies in the range $[-1, 2^k - 1]$.*

If $\textsc{Skew}_\psi(C) < 0$, we say that $C$ is negatively skewed while if $\textsc{Skew}_\psi(C) > 0$ we say that it is positively skewed. An averaging argument shows that the existence of negatively skewed subcubes implies the existence of positively skewed subcubes and vice versa.

▶ **Lemma 10.** *For any $K \subseteq [n]$, we have*

$$\sum_{\substack{D=(K,w) \\ w\in\{\pm 1\}^K}} \textsc{Skew}(D) = 0.$$

Given a cube $C = (K, y)$ of codimension $k$, we can partition it into $2^\ell$ subcubes of codimension $k + \ell$, where we pick a set $L$ of $\ell$ additional coordinates outside of $K$ to fix and enumerate over all settings of these coordinates.

▶ **Lemma 11.** *If $\{C_1, \ldots, C_{2^\ell}\}$ is a partition of $C$, then*

$$\textsc{Skew}_\psi(C) = \frac{1}{2^\ell} \sum_{i=1}^{2^\ell} \textsc{Skew}_\psi(C_i).$$

Having established basic properties of the skew function, we next turn to bounding the number of subcubes with a given skew. We show that this number may be quite large in the worst case.

▶ **Lemma 12.** *Let $\gamma = 2^f - 1$ for $f \in \{1, \ldots, k\}$. There exists a distribution $\psi$ such that there are $\Omega((n/k)^k)$ many subcubes of codimension $k$ with $\textsc{Skew}(C) \geq \gamma$.*

**Proof.** Let $C$ be the subcube where the first $t \geq f$ bits are fixed to 1, and let $\mu_C$ be the uniform distribution over it. Consider any subcube $D$ where we choose $f$ indices from $[t]$ and $k - f$ indices from $[n] \setminus [t]$, and set them to 1. We have

$$\langle \mu_C, \mu_D \rangle = \mathop{\mathbb{E}}_{\mu_C}[\mu_D] = 2^k \Pr_{\mathbf{x}\sim\mu_C}[\mathbf{x} \in D] = 2^k \frac{1}{2^{k-f}} = 2^f \geq 1 + \gamma$$

since a point from $\mu_C$ lies in $D$ iff the $k - f$ bits from $[n] \setminus [t]$ are all set to 1. We now optimize the choice of $t$. Let $\alpha = f/k$ for $\alpha \leq 1$. We choose $t = \alpha n$ (ignoring floors and ceilings which will not affect the asymptotics). The number of choices for $D$ is given by

$$\binom{t}{f} \cdot \binom{n-t}{k-f} = \binom{\alpha n}{\alpha k} \cdot \binom{(1-\alpha)n}{(1-\alpha)k} \geq \left(\frac{n}{k}\right)^{\alpha k} \cdot \left(\frac{n}{k}\right)^{(1-\alpha)k} \geq \left(\frac{n}{k}\right)^k. \qquad \blacktriangleleft$$

While the above bound is proved for positive skew, Lemma 10 can be used to derive a similar bound for negative skew. Given that this bound is not too far from the trivial upper bound of $\binom{n}{k}$, we need to refine our notion of skew, and also to restrict the set of distributions we consider.

## 3.2 Minimal skewed subcubes

Lemma 11 tells us that if there exists $C = (J, y)$ such that $|J| = j < k$ and $\textsc{Skew}(C) \geq \gamma$, then for any $L \subseteq [n] \setminus J$ of size $k - j$, there exists some further restriction of bits in $L$ such that the resulting subcube $D \subseteq C$ has $\textsc{Skew}(D) \geq \gamma$. This suggests that we ought to ignore subcubes such as $D$ that can be viewed as *inheriting* skew from some parent $C$, and instead focus on subcubes whose skew is larger than any parent. One technical issue is that we now need to handle the case of positive and negative skew separately. This motivates the following definitions.

▶ **Definition 13.** *Let $\gamma \in (0, 2^k - 1]$ and $\epsilon \in (0, 1]$. A subcube $C \subseteq \{\pm 1\}^n$ is a $(\gamma, \epsilon)$-minimally skewed subcube if $\textsc{Skew}(C) \geq \gamma$ and for all its parent subcubes $D \supsetneq C$, we have*

$$\textsc{Skew}_\psi(D) \leq (1 - \epsilon)\gamma. \tag{2}$$

*Let $\gamma \in (0, 1]$ and $\epsilon \in (0, 1]$. A subcube $C \subseteq \{\pm 1\}^n$ is a $(-\gamma, \epsilon)$-minimally skewed subcube if $\textsc{Skew}(C) \leq -\gamma$ and for all its parent subcubes $D \supsetneq C$, we have*

$$\textsc{Skew}_\psi(D) \geq -(1 - \epsilon)\gamma. \tag{3}$$

Note that our convention is to always use $\gamma > 0$ for the magnitude of the skew, and specify its sign explicitly. Note that the allowable values of $\gamma$ are different for the case of positive and negative skew. We restrict $\epsilon \in (0, 1]$. The case $\epsilon = 1$ corresponds to the case where every subcube of $C$ has no skew.

The crux of this definition is that minimal skew cannot be inherited from a parent. Given a minimal skewed subcube $C$, and a parent $D \supsetneq C$, we show that $C$ has noticeable skew in the restriction $\psi|_D$.

▶ **Lemma 14.** *If $C$ is a $(\gamma, \epsilon)$-minimal skewed subcube and $D \supseteq C$ is a parent of $C$, then*

$$\textsc{Skew}_{\psi|_D}(C) \geq \frac{\epsilon\sqrt{\gamma}}{2}.$$

*If $C$ is a $(-\gamma, \epsilon)$-minimal skewed subcube and $D \supseteq C$ is a parent of $C$, then*

$$\textsc{Skew}_{\psi|_D}(C) \leq -\epsilon\gamma.$$

**Proof.** We first consider the case when $\gamma > 0$. By Lemma 6

$$\langle \psi|_D , \mu_C|_D \rangle = \frac{\langle \psi, \mu_C \rangle}{\langle \psi, \mu_D \rangle} = \frac{1 + \textsc{Skew}_\psi(C)}{1 + \textsc{Skew}_\psi(D)} \geq \frac{1 + \gamma}{1 + (1 - \epsilon)\gamma}. \tag{4}$$

We have

$$\textsc{Skew}_{\psi|_D}(C) = \langle \psi|_D , \mu_C|_D \rangle - 1 \geq \frac{\epsilon\gamma}{1 + (1 - \epsilon)\gamma} \geq \frac{\epsilon\gamma}{2\sqrt{(1 - \epsilon)\gamma}} \geq \frac{\epsilon\sqrt{\gamma}}{2}$$

where the first inequality is by Equation (4) and the second is by the AM-GM inequality.

Next we consider the case where $\gamma < 0$. By Lemma 6

$$\langle \psi|_D , \mu_C|_D \rangle = \frac{\langle \psi, \mu_C \rangle}{\langle \psi, \mu_D \rangle} = \frac{1 + \textsc{Skew}_\psi(C)}{1 + \textsc{Skew}_\psi(D)} \leq \frac{1-\gamma}{1-(1-\epsilon)\gamma}.$$

Hence

$$\textsc{Skew}_{\psi|_D}(C) = \langle \psi|_D , \mu_C|_D \rangle - 1 \leq \frac{1-\gamma}{1-(1-\epsilon)\gamma} - 1 = \frac{\epsilon\gamma}{1-(1-\epsilon)\gamma} \leq -\epsilon\gamma. \quad \blacktriangleleft$$

## 4 Fourier Analysis

Given $S \subseteq [n]$, let $\chi_S : \{\pm 1\}^n \to \{\pm 1\}$ be given by $\chi_S(x) = \prod_{i \in S} x_i$. These functions form a basis so we can write $\psi = \sum_S \widehat{\psi}(S)\chi_S$, where the Fourier coefficients of $\psi$ are given by

$$\widehat{\psi}(S) = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mu}[\psi(\mathbf{x})\chi_S(\mathbf{x})] = \sum_{x \in \{\pm 1\}^n} \frac{\psi(x)\chi_S(x)}{2^n} = \sum_{x \in \{\pm 1\}^n} \Pr_{\mathbf{x}\sim\psi}[\mathbf{x} = x]\chi_S(x) = \mathop{\mathbb{E}}_{\mathbf{x}\sim\psi}[\chi_S(\mathbf{x})]$$

which is simply the bias of $\chi_S$ under the distribution $\psi$. Thus we have

$$\psi(x) = \sum_{S \subseteq [n]} \widehat{\psi}(S)\chi_S(x)$$

where $\widehat{\psi}(\emptyset) = \mathbb{E}_{\mathbf{x}\sim\psi}[\psi(x)] = 1$. Given two distributions $\psi$ and $\omega$, their inner product is given by

$$\langle \psi, \omega \rangle = \mathop{\mathbb{E}}_{\mathbf{x}\sim\{\pm 1\}^n}[\psi(\mathbf{x})\omega(\mathbf{x})] = \sum_{x \in \{\pm 1\}^n} \frac{\psi(x)\omega(x)}{2^n} = 1 + \sum_{\emptyset \neq S \subseteq [n]} \widehat{\psi}(S)\widehat{\omega}(S)$$

### Skew implies heavy low-degree coefficients

We show that large skew in the subcube $(K, y)$ implies non-trivial Fourier mass on subsets of $K$.

▶ **Lemma 15.** *For $C = (K, y)$,*

$$\textsc{Skew}_\psi(C) = \sum_{\emptyset \neq S \subseteq K} \widehat{\psi}(S)\chi_S(y).$$

**Proof.** Given $C = (K, y)$, $\mu_C$ the uniform measure over $C$ is given by

$$\mu_C(x) = \prod_{i \in K}(1 + x_i y_i) = 1 + \sum_{\emptyset \neq S \subseteq K} \chi_S(y)\chi_S(x).$$

Hence we have

$$\langle \psi, \mu_C \rangle = 1 + \sum_{\emptyset \neq S \subseteq [n]} \widehat{\psi}(S)\widehat{\mu}_A(S) = 1 + \sum_{\emptyset \neq S \subseteq K} \widehat{\psi}(S)\chi_S(y)$$

from which the claim follows. ◀

Given the above lemma, our approach is to reduce bounding the number of skewed subcubes to bounding the number of large Fourier coefficients of $\psi$ at level $k$.

We define

$$W^{\leq k}(\psi) = \sum_{\substack{S\subseteq[n]\\|S|\leq k}} \widehat{\psi}(S)^2.$$

A trivial bound is obtained from Parseval's identity:

$$W^{\leq k}(\psi) \leq \sum_{S\subseteq[n]} \widehat{\psi}(S)^2 = \mathop{\mathbb{E}}_{\mathbf{x}\sim\{\pm1\}^n}[\psi(\mathbf{x})^2] \leq \|\psi\|_\infty \mathop{\mathbb{E}}_{\mathbf{x}\sim\{\pm1\}^n}[\psi(\mathbf{x})] = \|\psi\|_\infty$$

If we restrict the summation to sets $S$ of cardinality at most $k$, then a much stronger bound of $O(\ln(\|\psi\|_\infty)^k)$ holds, it is proved using the powerful HyperContractivity Theorem. These bounds generalize the Level-k inequalities for the Fourier spectrum of small-sets, indeed the proof is identical.

For a $f : \{\pm1\}^n \to \mathbb{R}$ and $0 \leq \rho \leq 1$ set

$$T_\rho f = \sum_S \rho^{|S|} \widehat{f}(S)\chi_S$$

$T_\rho$ is known as the noise operator. Recall that for $p > 0$ we have $\|f\|_p = \mathbb{E}[f^p]^{1/p}$. The hypercontractive inequality quantifies the extent to which the noise operater reduces the norm of a function. See for instance. [16, Chapter 2] for a detailed exposition.

▶ **Theorem 16.** *Let $f : \{\pm1\}^n \to \mathbb{R}$ and $\rho \in [0, 1]$. Then*

$$\|T_\rho f\|_2 \leq \|f\|_{1+\rho^2}.$$

We use the hypercontractive inequality to bound the mass of the low level coefficients.

▶ **Theorem 17.** *Let $\psi$ be a distribution. Then*

$$W^{\leq k}(\psi) \leq e^2(\ln(e\,\|\psi\|_\infty))^k.$$

**Proof.** By Theorem 16, we have

$$\begin{aligned}
\|T_\rho\psi\|_2 &\leq \|\psi\|_{1+\rho^2} \\
&= \left(\mathop{\mathbb{E}}_{\mathbf{x}\sim\{\pm1\}^n}[\psi(x)^{\rho^2}\psi(x)]\right)^{1/(1+\rho^2)} \\
&\leq \|\psi\|_\infty^{\rho^2/(1+\rho^2)} \mathop{\mathbb{E}}_{\mathbf{x}\sim\{\pm1\}^n}[\psi(x)]^{1/(1+\rho^2)} \\
&= \exp\left(\frac{\ln(\|\psi\|_\infty)\rho^2}{1+\rho^2}\right)
\end{aligned}$$

where we used Holder's inequality with $p = \infty$, $q = 1$. Taking $\rho = \min(1, 1/\sqrt{\ln(\|\psi\|_\infty)})$, we have

$$\|T_\rho\psi\|_2 \leq \exp(1/(1+\rho^2)) \leq e$$

But note that

$$\|T_\rho\psi\|_2 \geq (\rho^{2k} W^{\leq k}(\psi))^{1/2}$$

Hence we conclude that

$$W^{\leq k}(\psi)) \leq e^2(1/\rho)^{2k} = e^2 \max(1, \ln\|\psi\|_\infty)^k \leq e^2(\ln(e\,\|\psi\|_\infty))^k.$$ ◀

We also need a bound for the Fourier mass at level $k$ where we do not count coordinates from some set $J \subseteq [n]$ in the degree of a coefficient.

$$W^{\leq k}(\psi, J) = \sum_{T \subseteq J} \sum_{\substack{S \subseteq [n] \setminus J \\ |S| \leq k}} \widehat{\psi}(S \cup T)^2.$$

▶ **Corollary 18.** *For $J \subseteq [n]$ and a distribution $\psi$ over $\{\pm 1\}^n$,*

$$W^{\leq k}(\psi, J) \leq 2^{|J|} e^2 (\ln(e \|\psi\|_\infty))^k.$$

## Projection, Extension, Restriction

Given $x \in \{\pm 1\}^n$ and a set of coordinates $P \subseteq [n]$, let $x_P$ denote the projection of $x$ onto coordinates in $P$. Given a distribution $\psi$ over $\{\pm 1\}^n$ and a set of coordinates $P \subseteq [n]$, let $\psi_P$ denote the marginal distribution over the set $P$. The Fourier expansion is especially convenient for marginals, we simply restrict the sum to subsets of $P$.

▶ **Lemma 19.** *For $P \subseteq [n]$ and a distribution $\psi$ over $\{\pm 1\}^n$, the restriction $\psi_P$ is given by*

$$\psi_P(y) = \sum_{S \subseteq P} \widehat{\psi}(S) \chi_S(y).$$

Coversely we can *extend* a distribution $\psi'$ defined on $\{\pm 1\}^P$ for $P \subseteq [n]$ to all of $\{\pm 1\}^n$ while preserving its important properties.

▶ **Lemma 20.** *Let $P \subsetneq [n]$. Let $\psi'$ be a distribution on $\{\pm 1\}^P$. Define a distribution $\psi$ on $\{\pm 1\}^n$ by $\psi(x) = \psi'(x_P)$. Then*

1. *$\psi$ is the product distribution of $\psi'$ with the uniform distribution on $\{\pm 1\}^{\bar{P}}$.*
2. *$\|\psi\|_\infty = \|\psi'\|_\infty$.*
3. *$C$ is a minimal skewed subcube under $\psi$ iff it is a minimal skewed subcube under $\psi'$.*

Finally, we derive an expression for the Fourier expansion of $\psi|_C$ in terms of the coefficients of $\psi$.

▶ **Lemma 21.** *Let $C = (J, z)$. Then*

$$\psi|_C(x) = \sum_{S \subseteq [n] \setminus J} \chi_S(x) \frac{\sum_{T \subseteq J} \widehat{\psi}(S \cup T) \chi_T(z)}{\langle \psi, \mu_C \rangle}.$$

## 5 A Combinatorial Bound for minimal skewed subcubes

In this section, we show bounds on the number of minimal skewed subcubes that is dimension independent.

▶ **Theorem 22** (Combinatorial Bound for Positive Skew). *For any measure $\psi$ on $\{\pm 1\}^n$, integer $k \leq n$, and $\gamma \in (0, 2^k - 1]$ and $\epsilon \in (0, 1]$, the number of $(\gamma, \epsilon)$-minimal skewed subcubes of codimension at most $k$ is bounded by*

$$k^{O(k)} \left( \frac{1}{\epsilon^2 \gamma} \ln(e \|\psi\|_\infty) \right)^k.$$

▶ **Theorem 23** (Combinatorial Bound for Negative Skew). *For any measure $\psi$ on $\{\pm 1\}^n$, integer $k \leq n$, and $\gamma \in (0,1]$ and $\epsilon \in (0,1]$, the number of $(-\gamma,\epsilon)$-minimal skewed subcubes of codimension at most $k$ is bounded by*

$$k^{O(k)} \left( \frac{1}{\epsilon^2 \gamma^2} \ln \left( \frac{e \, \|\psi\|_\infty}{\epsilon \gamma} \right) \right)^k .$$

We now outline our approach for proving these bounds.

1. We give an algorithm to enumerate all minimal skewed subcubes, given the list of large, low-degree Fourier coefficients in an adaptively chosen sequence of restrictions of the original distribution $\psi$. The algorithm recursively "grows" skewed subcubes by finding heavy Fourier coefficients and restricting the bits in that coefficient, and showing that this algorithm discovers all minimal skewed subcubes.
2. We bound the number of large low-degree Fourier coefficients of $\psi$ using Theorem 17.

The details of the algorithm in Step 1 are different for the cases of positive and negative skew, so we present them in Subsections 5.1 and 5.2. To go from a combinatorial bound to an efficient algorithm, we need to make Step 2 algorithmic. We will consider this problem under different learning models in Sections 6 and 8.

## 5.1 Positive skew

We first present an algorithm FindSkew$^+$ for enumerating minimal skewed subcubes where the skew is positive.

To prove the combinatorial bound, we allow the algorithm to make certain *guesses* in Lines 6 and 7. We think of the set of all possible outputs over all possible guesses as the list that is returned by the algorithm. In Lemma 24, we will show that all minimal skewed subcubes are contained in this list. We bound the list size in Lemma 26. Together, these complete the proof of Theorem 22.

We start the recursion with $R_0 = \emptyset$ and $z_0 = \emptyset$ the null string. The routine either returns FAIL or returns $S_t \subset [n]$ and $z_t \in \{\pm 1\}^{S_t}$ such that $(R_t, z_t)$ is a $\gamma$-skewed subcube. The algorithm also takes as inputs the input the distribution $\psi$, a bound $k$ on the codimension, and skew parameters $\gamma \in (0, 2^k - 1]$ and $\epsilon \in (0,1]$. These stay constant through the recursion, so we suppress the dependence on them. Consider the list of all possible choices returned by the algorithm.

**Algorithm 1** FindSkew$^+(R_t, y_t)$.

1: Let $D_t = (R_t, y_t)$. Let $\psi_t = \psi|_{D_t}$. Let $k_t = k - |R_t|$.
2: **if** $\text{Skew}_\psi(D_t) > \gamma(1-\epsilon)$ **then**
3: **return** $D_t$
4: **if** $\langle \psi, \mu_{D_t} \rangle < (1+\gamma) \cdot 2^{-k_t}$ **then**
5: **return** FAIL
6: Pick $S_t$ such that $|S_t| \leq k_t$ and

$$|\widehat{\psi_t}(S_t)| \geq \frac{\epsilon\sqrt{\gamma}}{k_t \cdot \binom{k_t}{|S_t|}}. \tag{5}$$

7: Pick $z_t \in \{\pm 1\}^{S_t}$.
8: **return** FindSkew$^+(R_t \cup S_t, y_t \circ z_t)$.

We need some notation for the analysis. Let the sequence of subcubes produced by the algorithm be $D_0 \supsetneq D_1 \cdots \supsetneq D_\ell$. Let $s_t = |S_t|$.

▶ **Lemma 24.** *For every $(\gamma, \epsilon)$-minimal skewed subcube $C$ with $\mathrm{codim}(C) \le k$ there are sequences of choices of $S_t$ and $z_t$ (in Lines 6 and 7) so that $C$ is returned by* FindSkew$^+$.

**Proof.** For every $C = (K^*, z^*)$ that is a $(\gamma, \epsilon)$-minimal skewed subcube where $\mathrm{codim}(C) \le k$, we will show that for every $t$, if $D_t \supsetneq C$ is parent of $C$, and is not equal to $C$ there is a choice of $S_t, z_t$ that leads to a parent $D_{t+1}$ of $C$ with a larger codimension. Since $t \le \mathrm{codim}(D_t) \le \mathrm{codim}(C) \le k$, in $\ell \le k$ steps we must have $D_t = C$, at which point we return at Line 3. Thus the claim implies the lemma.

At $t = 0$, we have $D_0 = \{\pm 1\}^n$ so the parent condition holds trivially. Assume that we have $D_t \supsetneq C$.

By the definition of a minimal skewed subcube, $\textsc{Skew}(D_t) \le (1-\epsilon)\gamma$, hence the procedure will not return at Line 3.

Next we show that $\langle \psi, \mu_{D_t} \rangle \ge (1+\gamma)2^{-k_t}$, the algorithm will not return FAIL at Line 5:

$$\begin{aligned}
\langle \psi, \mu_{D_t} \rangle &= \Pr_{\mathbf{x}\sim\psi}[\mu_{D_t}(\mathbf{x})] = 2^{k-k_t} \Pr_{\mathbf{x}\sim\psi}[\mathbf{x} \in D_t] \\
&\ge 2^{k-k_t} \Pr_{\mathbf{x}\sim\psi}[\mathbf{x} \in C] \\
&\ge 2^{-k_t} \langle \psi, \mu_{C_t} \rangle \\
&\ge (1+\gamma)2^{-k_t}.
\end{aligned}$$

The first inequality holds because $D_t \supset C$, the second because $\mathrm{codim}(C) \le k$ and the last because we assume that $\textsc{Skew}(C) \ge \gamma$.

Recall that $\psi_t = \psi|_{D_t}$, and let $K_t = K \setminus S_t$. By Lemma 15,

$$\textsc{Skew}_{\psi_t}(C) = \sum_{\emptyset \ne S \subseteq K_t} \widehat{\psi_t}(S)\chi_S(y) = \sum_{k'} \sum_{\substack{\emptyset \ne S \subseteq K_t \\ |S|=k'}} \widehat{\psi_t}(S)\chi_S(y)$$

By Lemma 14, $\textsc{Skew}_{\psi_t}(C) \ge \epsilon\sqrt{\gamma}/2$ which implies that for some $k' \le k_t$, we have

$$\sum_{\substack{\emptyset \ne S \subseteq K_t \\ |S|=k'}} \widehat{\psi_t}(S)\chi_S(y) \ge \epsilon\sqrt{\gamma}/k_t$$

which in turn implies that for at least one $\emptyset \ne S_t \subseteq K_t$, we have

$$|\widehat{\psi_t}(S_t)| \ge \epsilon\sqrt{\gamma} / \left(k_t \cdot \binom{k_t}{k'}\right).$$

Assume that we pick this $S_t$ in Line 6 and $z_t = z^*|_S$ in Line 7. This ensures that $D_{t+1}$ is a parent of $C$ of larger codimension. ◀

We next bound the number of all possible outputs of the algorithm. The crux of the argument is to bound the number of large low-degree Fourier coefficients using Theorem 17. This in turn requires a bound on the infinity norm of $\psi_t$ which comes from passing the test in Line 4.

▶ **Lemma 25.** *The number of choices for $S_t$ satisfying Equation* (5) *is bounded by*

$$\frac{e^2}{\epsilon^2\gamma}(\ln(2^{k_t} e \|\psi\|_\infty))^{s_t} k_t^{4s_t+2}.$$

**Proof.** We bound $\|\psi_t\|_\infty$ as

$$\|\psi_t\|_\infty = \left\|\psi|_{D_t}\right\|_\infty \leq \frac{\|\psi\|_\infty}{\langle\psi, \mu_{D_t}\rangle} \leq \frac{\|\psi\|_\infty}{(1+\gamma)2^{-k_t}} \leq \|\psi\|_\infty 2^{k_t}$$

where the first inequality is from Fact 5 and we have $\langle\psi, \mu_{D_t}\rangle \geq (1+\gamma)2^{-k_t}$ since we check for this condition in Line 5. We now use Theorem 17 which gives

$$W^{\leq s_t}(\psi) \leq e^2(\ln(e\|\psi\|_\infty \cdot 2^{k_t}))^{s_t}.$$

Hence the number of choices for $S_t$ satisfying (5) is bounded by

$$W^{\leq s_t}(\psi_t)\left(\frac{k_t\binom{k_t}{s_t}}{\epsilon\sqrt{\gamma}}\right)^2 \leq \frac{e^2}{\epsilon^2\gamma}(\ln(2^{k_t}e\|\psi\|_\infty))^{s_t}k_t^{4s_t+2}.$$ ◀

▶ **Lemma 26.** *The total number of subcubes of codimension $k$ output by* FindSkew$^+$ *is at most:*

$$k^{O(k)}\left(\frac{\ln(e\|\psi\|_\infty)}{\epsilon^2\gamma}\right)^k.$$

**Proof.** Since $\sum_{t\leq\ell} s_t = k$, the sequence $\{s_t\}_{t=1}^{\ell}$ is a partition of $k$, and there are at most $k^k$ of them. Let us fix the sequence. The number of choices for $S_t$ is bounded by Lemma 25. Since $z_t \in \{\pm 1\}^{S_t}$, the number of choices for $z$ is $2^{s_t}$. Taking the product over all $t$, the number of possible outputs for FindSkew$^+$ is bounded by

$$\prod_{t=1}^{\ell}\frac{e^2}{\epsilon^2\gamma}(\ln(2^{k_t}e\|\psi\|_\infty))^{s_t}k_t^{4s_t+2}\cdot 2^{s_t}$$

We can bound

$$\prod_{t=1}^{\ell}(\ln(2^{k_t}e\|\psi\|_\infty))^{s_t} \leq \ln(2^k e\|\psi\|_\infty)^k \leq (k+\ln(e\|\psi\|_\infty))^k \leq (2k)^k(\ln(e\|\psi\|_\infty))^k.$$

$$\prod_{t=1}^{\ell}k_t^{4s_t+2}2^{s_t} \leq k^{5k+2}.$$

Including the $k^k$ choices for $s_1, \ldots, s_t$, the output list size is bounded by

$$\left(\frac{e^2}{\epsilon^2\gamma}\right)^k(\ln(e\|\psi\|_\infty))^k k^{7k+2} = k^{O(k)}\left(\frac{\ln(e\|\psi\|_\infty)}{\epsilon^2\gamma}\right)^k.$$ ◀

Together Lemma 24 and Lemma 26 complete the proof of Theorem 22.

## 5.2 Negative Skew

We now present an algorithm FindSkew$^-$ for the negative skewed case. The algorithm takes as input $\gamma \in (0,1]$ and $[\epsilon \in (0,1]$ and the goal is to list all $(-\gamma,\epsilon)$-minimal negatively skewed subcubes.

The main differences from FindSkew$^+$ are that once the skew is less than $-\gamma(1-\epsilon)$, we can return. Thus we can combine the Return statement (Line 3, and the the check in Line 5. Also, the bound on the coefficient size in Equation(6) now reflects the bound for the negative skew case in Lemma 14.

We have the following claim about the correctness of FindSkew$^-$.

**Algorithm 2** FindSkew$^-(R_t, y_t)$.

1: Let $D_t = (S_t, z_t)$. Let $\psi_t = \psi|_{D_t}$. Let $k_t = k - |S_t|$.
2: **if** $\mathrm{Skew}_\psi(D_t) < -\gamma(1-\epsilon)$ **then**
3: **return** $D_t$
4: Pick $S_t$ such that $|S_t| \le k_t$ and

$$|\widehat{\psi_t}(S_t)| \ge \frac{\epsilon\gamma}{k_t \cdot \binom{k_t}{|S_t|}}. \tag{6}$$

5: Pick $z_t \in \{\pm 1\}^{S_t}$.
6: **return** FindSkew$^-(R_t \cup S_t, y_t \circ z_t)$.

▶ **Lemma 27.** *For every $(\gamma,\epsilon)$-minimal skewed subcube $C$ with $\mathrm{codim}(C) \le k$ there are choices of subsets $S_t$ and $z_t$ (in Lines 6 and 7) so that $C$ is returned by* FindSkew$^-$.

We prove this by showing that for every $t$, if $D_t \supsetneq C$ is parent of $C$, and is not equal to $C$ there is a choice of $S_t, z_t$ that gives a parent $D_{t+1}$ of $C$ with a larger codimension. Indeed, we know that for any parent of $C$, inner product $\langle \psi, \mu_{D_t}\rangle$ is large enough to pass the test in Line 3. The rest of the proof is identical to that of Lemma 24 for the case of positive skew, so we do not repeat it.

The crux of the proof is to bound the number of choices for $S_t$ satisfying Equation (6).

▶ **Lemma 28.** *The number of choices for $S_t$ satisfying Equation* (6) *is bounded by*

$$\frac{e^2}{\epsilon^2\gamma^2}\left(\ln\left(\frac{e\,\|\psi\|_\infty}{\epsilon\gamma}\right)\right)^{s_t} k^{2s_t+2}.$$

**Proof.** To pass Line 3, it must hold that $\mathrm{Skew}_\psi(D_t) \ge -\gamma(1-\epsilon)$, hence

$$\langle \psi, \mu_{D_t}\rangle \ge 1 - \gamma(1-\epsilon) \ge \epsilon\gamma$$

since $\gamma \le 1$. So we bound $\|\psi_t\|_\infty$ as

$$\|\psi_t\|_\infty = \left\|\psi|_{D_t}\right\|_\infty = \frac{\|\psi\|_\infty}{\langle \psi, \mu_{D_t}\rangle} \le \frac{\|\psi\|_\infty}{\gamma\epsilon}.$$

Using Theorem 17 gives

$$W^{\le s_t}(\psi_t) \le e^2\left(\ln\left(\frac{e\,\|\psi\|_\infty}{\epsilon\gamma}\right)\right)^{s_t}.$$

Hence the number of choices for $S_t$ satisfying Equation (6) is bounded by

$$\begin{aligned}\frac{W^{\le s_t}\cdot\left(k_t\cdot\binom{k_t}{|S_t|}\right)^2}{\epsilon^2\gamma^2} &\le e^2\left(\ln\left(\frac{e\,\|\psi\|_\infty}{\epsilon\gamma}\right)\right)^{s_t}\frac{\left(k_t\cdot\binom{k_t}{|S_t|}\right)^2}{\epsilon^2\gamma^2}\\ &\le \frac{e^2}{\epsilon^2\gamma^2}\left(\ln\left(\frac{e\,\|\psi\|_\infty}{\epsilon\gamma}\right)\right)^{s_t} k^{2s_t+2}.\end{aligned}$$

◀

We can now conclude as before.

▶ **Lemma 29.** *The total number of subcubes of codimension $k$ output by* FindSkew$^-$ *is bounded by*

$$k^{O(k)}\left(\frac{1}{\epsilon^2\gamma^2}\ln\left(\frac{e\,\|\psi\|_\infty}{\epsilon\gamma}\right)\right)^k.$$

**Proof.** Using Lemma 28 and the fact that there are are $2^{s_t}$ choices for $z_t$ and $k^k$ choices of the partition $s_1, \ldots, s_t$, the overall list size is bounded by

$$k^k \prod_{t=1}^{\ell} \frac{e^2}{\epsilon^2\gamma^2} \left( \ln\left(\frac{e\,\|\psi\|_\infty}{\epsilon\gamma}\right)\right)^{s_t} k^{2s_t+2}2^{s_t} \leq k^{O(k)} \left(\frac{1}{\epsilon^2\gamma^2} \ln\left(\frac{e\,\|\psi\|_\infty}{\epsilon\gamma}\right)\right)^k. \quad \blacktriangleleft$$

Combining Lemmas 27 and 29 completes the proof of Theorem 23.

## 5.3 Tightness of our bounds

We show that the dependence on $\|\psi\|_\infty$ in Theorem 22 is nearly optimal. To simplify our constructions, we will construct distributions on $n$ variables where $n = n(\|\psi\|_\infty, k)$. But one can then use Lemma 20 to extend the construction to all larger values of $n$.

▶ **Theorem 30.** *There exists a distribution $\mu_C$ on $\{\pm1\}^n$ which has $\Omega_k((\ln(\|\mu_C\|_\infty))^k)$ many $(2^k, 1/2)$-minimal skewed subcubes of codimension $k$.*

**Proof.** Let $C$ be the subcube where all the bits are fixed to 1, and let $\mu_C$ be the uniform distribution over it. It follows that $\|\mu_C\|_\infty = 2^n$ hence $\ln(\|\mu_C\|_\infty) = n$. We claim that all $\binom{n}{k}$ subcubes where a $k$ out of the first $t$ bits are fixed to 1 are $(2^k - 1, 1/2)$-minimal skewed subcubes. Fix one such cube $D$. We have

$$\textsc{Skew}_{\mu_C}(D) = 2^k \Pr_{\mathbf{x}\sim\mu_C}[\mathbf{x} \in D] - 1 = 2^k - 1.$$

Since the maximum skew of any subcube of codimension $k-1$ is at most $2^{k-1} - 1$, $D$ satisfies the definition of $(\gamma, \epsilon)$-minimal skew for $\gamma = 2^k - 1$ and $\epsilon$ such that $\gamma(1-\epsilon) \geq 2^{k-1} - 1$. In particular, we can take $\epsilon = 1/2$.

Thus the number of $(2^k - 1, 1/2)$-minimal skewed subcubes is $\binom{n}{k} = \Omega_k(n^k)$. The only dependence on $n$ in Theorem 22 comes from the $\|\mu_C\|_\infty$ since $\epsilon$ is a constant and $\gamma \leq 2^k$. Thus the number of cubes is $\Omega((\ln(\|\mu_C\|_\infty))^k)$. $\blacktriangleleft$

For Theorem 23 dealing with negative skew, we show a similar bound, though with a smaller value of $\epsilon = 1/k$. The distribution we use is derived from the Tribes function.

▶ **Theorem 31.** *There exists a distribution $\tau$ on $\{\pm1\}^n$ which has $\Omega_k((\ln(\|\tau\|_\infty))^k)$ many $(-1, 1/k)$-minimal skewed subcubes of codimension $k$.*

**Proof.** Let $n = tk$. We label the coordinates as $\{x_{i,j}\}_{i\in[k],j\in[t]}$. Consider the DNF formula

$$\text{Tribes}(x) = \bigvee_{i=1}^{k} \bigwedge_{j=1}^{t} x_{i,j}$$

We now define a distribution $\tau$ where we pick a $i^* \in [k]$ at random, set $x_{i^*,j} = 1$ for all $j \in [t]$ and set all the other variables randomly. Clearly the distribution $\tau$ is supported on the satisfying assignments of Tribes$(x)$. It is also easy to see that

$$\|\tau\|_\infty = \tau(1^{tk}) = 2^{tk} \sum_{i=1}^{k} \frac{1}{k} 2^{-(k-1)t} = 2^t.$$

Now consider the set of minimal 0 certificates of Tribes, which are subcubes where we pick a single variables from each term and set it to 0. There are $t^k$ such subcubes, fix one such subcube $C$. Clearly $\Pr_{\mathbf{x}\sim\tau}[\mathbf{x} \in C] = 0$, hence $\textsc{Skew}_\tau(C) = -1$. Now consider any parent subcube $D$ of $C$. Assume it has codimension $\ell < k$, and let $L \subset [k]$ denote the set of terms that it sets to 0. For $\mathbf{x} \sim \tau$ to lie in $L$, two events need to happen:

- $i^* \notin L$, which happens with probability $1 - \ell/k$.
- The variables in $L$ which are set to 0 in $D$ are also set to 0 by the random assignment, which happens with probability $2^{-\ell}$.

As these two events are independent, we have $\Pr_{\mathbf{x}\sim\tau}[\mathbf{x} \in D] = 2^{-\ell}(1 - \ell/k)$ hence

$$\textsc{Skew}_\tau(D) = 2^\ell \Pr_{\mathbf{x}\sim\tau}[\mathbf{x} \in D] - 1 = -\frac{\ell}{k}.$$

Thus the maximum skew of any parent $D$ is $-1 + 1/k$. Hence $C$ is $(-1, 1/k)$-minimally skewed.

As before we note that $\gamma = 1$ and $\epsilon = 1/k$, hence the only dependence on $t$ comes from $\log(\|\tau\|_\infty) = t$, which gives the claimed bound. ◀

Finally, we present a distribution that has a large number of $(-1, 1)$ and $(1, 1)$ minimally skewed subcubes. Recall that $\epsilon = 1$ means that every parent of the cube has skew 0.

The construction is based on (dual) BCH codes. We think of linear codes as subsets of $\mathbb{F}_2^n$ where $\mathbb{F}_2 = \{0, 1\}$ which we can identify with $\{\pm 1\}^n$ via the usual mapping $x \to (-1)^x$. For $x \in \mathbb{F}_2^n$ let the weight of $x$ denoted $\mathrm{wt}(x)$ be the number of 1s in $x$. Let $\mathrm{supp}(x) \subseteq [n]$ denote the set of cordinates where $x$ is non-zero. We will use the following fact about BCH codes communicated to us by Sergey Yekhanin [19].

▶ **Lemma 32** ([19]). *Let $d \geq 2$ be even and let $n + 1 = 2^l \geq d$. There exists a $\mathbb{F}_2$-linear code $\mathcal{C}_{BCH} \subseteq \{0, 1\}^n$ with minimum distance $d$, which contains $\Omega(n^{d/2+1})$ minimum weight codewords.*

▶ **Theorem 33.** *For any even $k \geq 2$ and large enough $n$, there exists a distribution $\psi_k$ on $\{\pm 1\}^n$ where the numbers of $(-1, 1)$-minimal skewed subcubes and $(1, 1)$-minimal skewed subcubes of codimension $k$ are both $\Omega_k((\log(\|\psi_k\|_\infty))^{k/2+1})$.*

**Proof.** Set $k = d$, and take $n$ as in Lemma 32. Let $\psi^k$ be the uniform distribution on the dual space to $\mathcal{C}_{BCH}$. Using standard facts about the Fourier expansion of a subspace, we can write

$$\psi^k = \sum_{\substack{c \in \mathcal{C}_{BCH} \\ S = \mathrm{supp}(c)}} \chi_S(x). \tag{7}$$

Since $\mathcal{C}_{BCH}$ has minimum distance $k$, $\psi^k$ is $(k-1)$-wise independent, so for any subcube $C$ where $\mathrm{codim}(C) \leq k - 1$, we have $\textsc{Skew}_{\psi^k}(C) = 0$. This relies on a standard construction of $k$-wise independent spaces from codes [2, Chapter 16], it can also be seen using Lemma 15 combined with Equation (7).

Fix $S \subset [n]$ to be the support of a minimum weight codeword in $\mathcal{C}_{BCH}$. By Lemma 19 and Equation (7), the projection of $\psi^k$ to coordinates in $S$ is given by $\psi^k_S(x) = 1 + \chi_S(x)$. Hence it is uniform over the $2^{k-1}$ settings $y \in \{\pm 1\}^S$ such that $\chi_S(y^+) = 1$.

For every such $y$ and $D^+ = (S, y)$, we have $\Pr_{\mathbf{x}\sim\psi^k}[\mathbf{x} \in D^+] = 2^{-(k-1)}$ hence

$$\textsc{Skew}_{\psi^k}(D^+) = 2^k \Pr_{\mathbf{x}\sim\psi^k}[\mathbf{x} \in D^+] - 1 = 1.$$

On the other hand, for every $y \in \{\pm 1\}^S$ such that $\chi_S(y) = -1$ and $D^- = (S, y)$, we have $\Pr_{\mathbf{x}\sim\psi^k}[\mathbf{x} \in D^-] = 0$ hence

$$\textsc{Skew}_{\psi^k}(D^-) = 2^k \Pr_{\mathbf{x}\sim\psi^k}[\mathbf{x} \in D^-] - 1 = -1.$$

Since every parent of $D^+$ has 0 skew, every such $D^+$ is a $(1, 1)$-minimal skewed subcube, and similary for every $D^-$.

Trivially, we have $\|\psi^k\|_\infty \le 2^n$, hence $\log(\|\psi^k\|_\infty) \le n$ (in fact it equals $n - O(\log(n))$. Since the number of minimal weight codewords is $\Omega_k(n^{k/2+1})$ by Lemma 32, and $\gamma, \epsilon$ are both 1, the number of codewords is $\Omega_k((\log(\|\psi^k\|_\infty))^{k/2+1})$. Hence the number of minimal skewed subcubes is as claimed. ◀

## 6 Algorithms for Finding Skewed Subcubes

In this section, we present an algorithm that find skewed subcubes efficiently in the random sample model, where we have access to random samples from $\psi$.

To make Algorithm 1 efficient, we need to replace the step of guessing $S$ (Line 6 in Algorithm 1, and Line 4 in Algorithm 2) with an algorithm to find large low degree Fourier coefficients[4]. We restate the problem below:

▶ **Problem 3** (Finding large low-degree biases). *Given a distribution $\psi$ on $\{\pm 1\}^n$, an integer $k$ and $\rho \ge 0$, find all $S \subseteq [n]$ such that $|S| \le k$ and*

$$\hat{\psi}(S) := \mathop{\mathbb{E}}_{\mathbf{x}\sim\psi}[\chi_S(x)] \ge \rho.$$

Our main result is the following pair of theorems.

▶ **Theorem 34** (Algorithm for Positive Skew). *Given sample access to a distribution $\psi$ on $\{\pm 1\}^n$, integer $k \le n$, and parameters $\gamma \in (0, 2^k - 1]$, $\epsilon \in (0, 1]$ and $\lambda \in [0, 1]$, there is an algorithm that returns all $(\gamma, \epsilon)$-minimal skewed subcubes of codimension at most $k$ in time:*

$$\tilde{O}\left(n^{k\left(\frac{\omega}{3-\lambda}\right)}\right) + k^{O(k)} \cdot (\ln(e\,\|\psi\|_\infty))^k \left(\frac{\tilde{O}(n^{k/3})}{(\epsilon\sqrt{\gamma})^{4/\lambda}} + \frac{\mathrm{poly}(n)}{(\epsilon\sqrt{\gamma})^{2k}}\right)$$

*where $\omega$ is the matrix multiplication exponent, and $\tilde{O}$ hides $\mathrm{poly}\log n$ factors.*

▶ **Theorem 35** (Algorithm for Negative Skew). *Given sample access to a distribution $\psi$ on $\{\pm 1\}^n$, integer $k \le n$, and parameters $\gamma \in (0, 1]$, $\epsilon \in (0, 1]$ and $\lambda \in [0, 1]$, there is an algorithm that returns all $(-\gamma, \epsilon)$-minimal skewed subcubes of codimension at most $k$ in time:*

$$\tilde{O}\left(n^{k\left(\frac{\omega}{3-\lambda}\right)}\right) + k^{O(k)} \cdot (\ln(e\,\|\psi\|_\infty))^k \left(\frac{\tilde{O}(n^{k/3})}{(\epsilon\gamma)^{4/\lambda}} + \frac{\mathrm{poly}(n)}{(\epsilon\gamma)^{2k}}\right)$$

*where $\omega$ is the matrix multiplication exponent, and $\tilde{O}$ hides $\mathrm{poly}\log n$ factors.*

Theorem 3 follows from using $(a)$, setting $\lambda = 0.01$ and $\omega \le 2.38$.

In both algorithms, we will find large low-degree biases using a breakthrough algorithm of [18] for detecting pairs of vectors that are highly correlated from a set of weakly correlated vectors. The algorithm was subsequently improved by [13]).

▶ **Theorem 36** ([13]). *Given two sets of vectors $V_1, V_2 \subseteq \{\pm 1\}^n$ for which there are at most $q$ pairs $(v_1, v_2) \in V_1 \times V_2$ with correlation larger than $\tau$, and a parameter $\rho \ge \tau^{1/\lambda}$ (for $\lambda \in [0, 1]$), there is an algorithm $\textsc{FindCorr}(V_1, V_2, \rho, \tau)$ that with high probability outputs all pairs $(v_1, v_2) \in V_1 \times V_2$ with correlation at least $\rho$. Furthermore, algorithm runs in time*

$$\tilde{O}\left(n^{\frac{2\omega}{3-\lambda}} + qdn^{\frac{2(1-\lambda)}{3-\lambda}}\right).$$

[4] We note that technically to implement the algorithm we also need to estimate $\langle \psi, \mu_C \rangle$ for every $C$ to an additive accuracy of $\min(\gamma, 2^{-k})$. If done via sampling, these only incur $2^{2k} k \log n/\gamma^2$ additional cost per call, and will be absorbed into our runtime bounds anyway.

The essence of the reduction is as follows. For each set $S \subseteq [n]$ less than $k/2$ we associate a vector $y_S$, for which each coordinate is a random sample of $\chi_S(x)$ where $x$ is drawn from $\psi$. If $Q, R \subseteq [n]$ are disjoint, the correlation coefficient $\mathbb{E}[y_Q \cdot y_R]/d$ is precisely the value of the Fourier coefficient $\widehat{\psi}(Q \cup R)$. Thus as long as the algorithm of [13] succeeds and is not overwhelmed by sample error, every $\rho$ correlated pair $(y_Q, y_R)$ corresponds to a Fourier coefficient $\widehat{\psi}(Q \cup R)$ of size less than $k$ and absolute value $\rho$. We now describe this more formally.

**Algorithm 3** FindFourierCoefficients$(\psi, k, \rho, \lambda)$.

$\mathcal{S} \leftarrow \emptyset$
$\tau \leftarrow (\rho/2)^{1/\lambda}$
Draw a set of $d = O(k \log n/\tau^2)$ samples $x_1, \ldots x_d$ from $\psi$.
**for** $T = O(k^{3/2} \log n)$ rounds **do**
  Randomly partition $[n]$ into two subsets $N_1$ and $N_2$.
  For every subset $S \subseteq N_1$ of size $\leq \lceil k/2 \rceil$, form a vector $y_S \in [-1, 1]^d$ for which the $i$th bit is set to $\chi_S(x_i)$. Call this set of vectors $V_1$.
  Do the same for $N_2$ for sets of size $\leq \lfloor k/2 \rfloor$, and call the set of vectors $V_2$.
  Run FindCorr$(V_1, V_2, \rho/2, \tau)$ from [13] to find all pairs $y_Q$ and $y_R$ such that $Q \subseteq N_1$, $R \subseteq N_2$, and $y_Q$ and $y_R$ are $\rho/2$ correlated. For each of these, add $Q \cup R$ to $\mathcal{S}$.
**return** $\mathcal{S}$.

We first prove some simple lemmas using standard concentration of measure results.

▶ **Lemma 37.** *For every $Q \in N_1$ and $R \in N_2$, w.h.p. $\left|\langle y_Q, y_R \rangle / d - \widehat{\psi}(Q \cup R)\right| \leq \tau/2$.*

**Proof.** For a single $x \sim \{\pm 1\}^n$ and disjoint $Q, R \subseteq N$, we have $\mathbb{E}[\chi_Q(x) \cdot \chi_R(x)] = \widehat{\psi}(Q \cup R)$. Applying the Hoeffding bound with our choice of $d = 32k \log n/\tau^2$, we have that $\langle y_Q, y_R \rangle / d$ is within $\tau/2$ of $\widehat{\psi}(Q \cup R)$ with probability at least $1 - n^{-2k}$. By a union bound, this hold for all $O(n^k)$ choices of pairs $(Q, R)$ with probability $1 - n^{-k}$. ◀

▶ **Lemma 38.** *For every $S \subseteq [n]$ with $|S| \leq k$, w.h.p. in at least one round $t \in [T]$ there are $Q \subseteq N_1$, $R \subseteq N_2$ with $Q \cap R = \emptyset$ such that $Q \cup R = S$.*

**Proof.** Fix a set $S$ of size $\ell$. For a random bipartition of $[n]$, the probability $S$ is perfectly bisected is at least $1/(8\sqrt{\ell}) \geq 1/(8\sqrt{k})$. The probability it is never bisected over $T = 16k^{3/2} \log n$ rounds is upper bounded by

$$\left(1 - \frac{1}{8\sqrt{k}}\right)^T \leq e^{-2k \log n} = n^{-2k}.$$

By a union bound, every $S$ of size $\leq k$ is bisected at least once with high probability. ◀

▶ **Lemma 39.** *The algorithm* FindFourierCoefficients$(\psi, k, \rho, \lambda)$ *returns all Fourier coefficients of $\psi$ of degree at most $k$ of absolute value at least $\rho$ in time*

$$\tilde{O}\left(n^{k\omega/(3-\lambda)}\right) + \tilde{O}(n^{k/3}) 2^{O(k)} (\ln(e \|\psi\|_\infty))^k \rho^{-4/\lambda}.$$

**Proof.** Consider any set $S \subseteq [n]$ of size $\leq k$ of magnitude at least $\rho$. By Lemma 38, w.h.p. for some round $t \in [T]$, the algorithm will form two vectors $y_Q$ and $y_R$ such that $Q \subseteq N_1$, $R \subseteq N_2$ and $Q \cup R = S$. Furthermore, by Lemma 37, we have $\langle y_Q, y_R \rangle / d \geq \rho - \tau \geq \rho/2$. In turn, this means that FindCorr$(V_1, V_2, \rho/2, \tau)$ will detect these w.h.p.

To bound the running time, we need a bound on the number $q$ of pairs with correlation higher than $\tau$. By Lemma 37, we have $\langle y_Q, y_R \rangle / d \geq \tau$ implies $\hat{\psi}(Q \cup R) \geq \tau/2$. The number of such coefficients is bounded by

$$\frac{W^{\leq k}(\psi)}{(\tau/2)^2} \leq 2^8 \left(\ln(e \|\psi\|_\infty)\right)^k \rho^{-2/\lambda}.$$

For each such coefficient $S$, there are $2^k$ ways to write it as $S = Q \cup R$ for disjoint $Q, R$. Hence

$$q \leq 2^{k+8} (\ln(e \|\psi\|_\infty))^k \rho^{-2/\lambda}.$$

Observe that $\log_\tau \rho \geq \lambda$ by our choice of $\tau$. By Theorem 36, [13] will find a list containing all $\rho/2$ correlated pairs in time at most

$$\begin{aligned}
&\tilde{O}\left(\left(n^{k/2}\right)^{2\omega/(3-\lambda)} + qd\left(n^{k/2}\right)^{(2-2\lambda)/(3-\lambda)}\right) \\
&\leq \tilde{O}\left(n^{k\omega/(3-\lambda)} + 2^{O(k)}(\ln(e\|\psi\|_\infty))^k \rho^{-2/\lambda} k \cdot \log(n) \cdot \rho^{-2/\lambda} n^{k(1-\lambda)/(3-\lambda)}\right) \\
&\leq \tilde{O}\left(n^{k\omega/(3-\lambda)}\right) + \tilde{O}(n^{k/3}) 2^{O(k)} (\ln(e\|\psi\|_\infty))^k \rho^{-4/\lambda}.
\end{aligned}$$

◀

We relegate the remainder of the proofs of Theorem 34 to the appendix.

## 7 Reduction from Noisy Parity

Recall that given $S \subseteq [n]$, a parity function $\chi_S : \{\pm 1\}^n \to \{\pm 1\}$ is given by $\chi_S(x) = \prod_{i \in S} x_i$. A noisy parity is a parity function with random noise of rate $\eta$ added to it. In other words, we say $f : \{\pm 1\}^n \to \{\pm 1\}$ is an $\eta$-noisy parity if $\Pr[f(x) = \chi_S(x)] = 1 - \eta$ an $\Pr[f(x) = -\chi_S(x)] = \eta$. In the sparse noisy parity problem, we are given access to samples $(\mathbf{x}, f(\mathbf{x}))$ where $\mathbf{x} \sim \mu$ is sampled uniformly from $\{\pm 1\}^n$ and $f$ is noisy parity $\chi_S$ with $|S| = k$. The goal is to recover the parity function, or equivalently the set $S$.

Given a set $S' \subseteq [n]$, we have

$$\mathop{\mathbb{E}}_{\mathbf{x} \sim \mu}[\widehat{f}(S')] = \begin{cases} 0 & \text{if } S' \neq S \\ 1 - 2\eta & \text{if } S' = S \end{cases}$$

This leads to a naive enumeration algorithm that runs in time $O(n^k)$. The current best algorithm due to [18] runs in time $O(n^{0.8k})\,\mathrm{poly}(1/(1-2\eta)))$. A series of reductions due to [8] show that efficient algorithms for sparse noisy parity imply algorithms with similar running times for learning $k$-juntas, decision trees and DNFs under the uniform distribution. This suggests the following conjecture (which we consider to be folklore).

▶ **Conjecture.** *There is no algorithm for the sparse noisy parity problem which runs in time* $n^{o(k)}$.

We now prove Theorem 2 which we restate below.

▶ **Theorem 2.** *For $\eta \in (0, 1/2)$, an algorithm that given a distribution $\psi$ and $k$ can find a $(1-2\eta, 1)$-minimal skewed subcube of co-dimension $k$ in time $T(n, k, \eta)$ can be used to solve the $k$-sparse noisy parity problem with noise rate $\eta$ in time $T(n, k, \eta)$.*

**Proof.** Finding a noisy parity reduces to finding a minimal skewed subcube. Given an instance of sparse noisy parity, consider the distribution $\psi$ on $\{\pm 1\}^{n+1}$ obtained by appending the label to the sample. Thus $\psi = (\mathbf{x}, f(\mathbf{x}))_{\mathbf{x}\sim\mu}$. We show that all skewed subcubes must restrict the set $S \cup \{n+1\}$, hence they are all minimal, and finding any skewed subcube solves the noisy parity problem.

For any $z \in \{\pm 1\}^S$, consider the subcube $D^+(z)$ given by $x_S = z$, $x_{n+1} = \chi_S(z)$. We have

$$\textsc{Skew}(D^+(z)) = 2^{k+1}\left(\Pr_{\mathbf{x}'\sim\psi}[\mathbf{x}' \in D] - \Pr_{\mathbf{x}'\sim\mu}[\mathbf{x}' \in D]\right) = 2^{k+1}\left(\frac{1-\eta}{2^k} - \frac{1}{2^{k+1}}\right) = 1 - 2\eta.$$

Similarly if we define $D^-(z)$ by $x_S = z$ and $x_{n+1} = -\chi_S(z)$, then

$$\textsc{Skew}(D^-(z)) = 2^{k+1}\left(\Pr_{\mathbf{x}'\sim\psi}[\mathbf{x}' \in D] - \Pr_{\mathbf{x}'\sim\mu}[\mathbf{x}' \in D]\right) = 2^{k+1}\left(\frac{\eta}{2^k} - \frac{1}{2^{k+1}}\right) = -(1-2\eta).$$

It is easy to verify that if the set of restricted variables is not $S \cup \{n+1\}$, then the skew is 0, which shows that the subcubes above all $(\pm(1-2\eta), 1)$-minimal skewed subcubes. ◀

This shows the hardness of finding subcubes with skew $(1-2\eta) < 1$. The reduction could be extended to show the hardness of finding subcubes with larger skew, simply by concatenating $\ell$ different samples of $\psi$. Now an algorithm that finds a subcube of skew $(2(1-\eta))^\ell - 1$ and codimension $k$ can be used to solve a $k/\ell$-sparse noisy parity problem.

## 8 The Membership Query Model

Theorem 2 suggests that a much better algorithm does not exist in the model where we only get random samples from $\psi$. However, noisy parity becomes trivial when we are given query access to $f$, by repeatedly querying the function at $x$ and $x \cdot e_i$. This motivates us to consider the query model where in addition to getting random samples from $\psi$, we are allowed to query $\psi(x)$ for points $x$ of our choosing. As we will see, this does make finding skewed subcubes easier for distributions where $\|\psi\|_\infty$ is small. We first show how this improvement arises, and then give evidence that queries do not add too much power over random samples when $\|\psi\|_\infty$ is large.

Algorithmically, all we need is a procedure to find all large low-degree Fourier coefficients of $\psi$ under the query model. Such a procedure is given by a classic result [14] of Kushilevitz and Mansour, which uses the algorithm of Goldreich and Levin [9] to compute large Fourier coefficients when given a query access to a function.

▶ **Theorem 40** ([9]). *Given query access to $f : \{\pm 1\}^n \to [-t, t]$ and a parameter $\rho > 0$, there is an algorithm running in time poly$(n, t/\rho)$ that with high probability outputs a list containing all subsets $S$, such that $\widehat{f}(S) \geq \tau$.* [5]

If we apply it to $\psi$, then we get an algorithm whose running time is poly$(n, \|\psi\|_\infty, \rho)$. Thus, the algorithm is faster than the trivial exhaustive search algorithm only when $\|\psi\|_\infty < n^{\alpha k}$ for some $\alpha > 0$. The polynomial dependence on $\|\psi\|_\infty$ in the running time is inevitable since the algorithm finds all $\widehat{\psi}(S) \geq \rho$, and not just those with $|S| \leq k$. The number of such coefficients can be $\|\psi\|_\infty / \rho^2$. In contrast, when we restrict to $|S| \leq k$, the list-size only grows as $\ln(e\|\psi\|_\infty)^k/\rho^2$. This raises the following natural open question:

[5] The theorem is typically stated for functions with range $\{\pm 1\}$, however a similar bound is true for the range $[-1, 1]$. The version stated here follows by scaling the function by $t$ so it lies in the range $[-1, 1]$, and replacing $\rho$ by $\rho/t$.

▶ **Problem 4.** *Given query access to a probability measure, $\psi$ such parameter $\rho > 0$, does there exist an algorithm that can find all $S$ such that $|S| \leq k$ and $|\widehat{\psi}(S)| \geq \rho$ in time* $\mathrm{poly}(n, 1/\rho, \ln(\|\psi\|_\infty)^k)$*?*

We conclude by observing that some dependence on $\|\psi\|_\infty$ (at least logarithmic) seems inevitable, even in cases where the list-size is 1. This is seen by a reduction from sparse noisy parity. A sample of size $O(k \log n/\epsilon^2)$ from an instance of noisy parity preserves all correlations of sets of $k$ variables up to an additive $\epsilon$. Define $\psi$ to be the uniform measure of these samples alone. Finding $S$ such that $|S| \leq k+1$ and $\widehat{\psi}(S) \geq 1 - 2\eta$ will solve the noisy parity problem. Note that for $\psi$ the query model and the random samples model are equivalent as we have the support explicitly. So if we believe that sparse noisy parity requires time $n^{\Omega(k)}$ time, then any algorithm for finding large low-degree Fourier coefficients in $\psi$ must require as much time. Since $\|\psi\|_\infty = \epsilon^2 2^n/k \log(n)$, we have $\ln(\|\psi\|_\infty) = n - O(\log(1/\epsilon)$. This is consistent with a dependence of $\ln(\|\psi\|_\infty)^{\Omega(k)}$.

## References


1. Noga Alon, Alexandr Andoni, Tali Kaufman, Kevin Matulef, Ronitt Rubinfeld, and Ning Xie. Testing k-wise and almost k-wise independence. In *STOC*, 2007.
2. Noga Alon and Joel H. Spencer. *The Probabilistic Method*. Wiley Publishing, 4th edition, 2016.
3. Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, pages 77–91, 2018. URL: `http://proceedings.mlr.press/v81/buolamwini18a.html`.
4. Ángel Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. FairVis: Visual Analytics for Discovering Intersectional Bias in Machine Learning. *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2019. URL: `https://poloclub.github.io/FairVis/`.
5. Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, 2009. `doi:10.1145/1541880.1541882`.
6. Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pages 47–60, 2017.
7. Andrew F Emmott, Shubhomoy Das, Thomas Dietterich, Alan Fern, and Weng-Keen Wong. Systematic construction of anomaly detection benchmarks from real data. In *Proceedings of the ACM SIGKDD workshop on outlier detection and description*, pages 16–21. ACM, 2013.
8. Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. On Agnostic Learning of Parities, Monomials, and Halfspaces. *SIAM J. Comput.*, 39(2):606–645, 2009. `doi:10.1137/070684914`.
9. Oded Goldreich and Leonid A. Levin. A Hard-Core Predicate for all One-Way Functions. In *Proceedings of the 21st Annual ACM Symposium on Theory of Computing, May 14-17, 1989, Seattle, Washigton, USA*, pages 25–32, 1989.
10. Parikshit Gopalan, Vatsal Sharan, and Udi Wieder. PIDForest: Anomaly detection via partial identification. In *Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, 2019.
11. Sudipto Guha, Nina Mishra, Gourav Roy, and Okke Schrijvers. Robust Random Cut Forest Based Anomaly Detection on Streams. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 2712–2721, 2016.
12. Sushrut Karmalkar, Adam R. Klivans, and Pravesh K. Kothari. List-Decodable Linear Regression. *CoRR*, abs/1905.05679, 2019. `arXiv:1905.05679`.

13 Matti Karppa, Petteri Kaski, and Jukka Kohonen. A Faster Subquadratic Algorithm for Finding Outlier Correlations. *ACM Trans. Algorithms*, 14(3):31:1–31:26, June 2018. doi:10.1145/3174804.

14 Eyal Kushilevitz and Yishay Mansour. Learning Decision Trees Using the Fourier Spectrum. *SIAM J. Comput.*, 22(6):1331–1348, 1993. doi:10.1137/0222080.

15 Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation Forest. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*, pages 413–422, 2008.

16 Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

17 Ryan O'Donnell and Yu Zhao. On Closeness to k-Wise Uniformity. In *APPROX-RANDOM*, 2018.

18 Gregory Valiant. Finding Correlations in Subquadratic Time, with Applications to Learning Parities and the Closest Pair Problem. *J. ACM*, 62(2):13:1–13:45, May 2015. doi:10.1145/2728167.

19 Sergey Yekhanin. Personal Communication, 2019.


## 9 Missing Proofs

▶ **Lemma 6.** *Given subcubes $C$ and $D$ such that $D \subseteq C \subseteq \{\pm 1\}^n$, and a density function $\psi$, it holds that:*

$$\langle \psi, \mu_D \rangle = \langle \psi, \mu_C \rangle \cdot \left\langle \psi|_C, \mu_D|_C \right\rangle$$

**Proof.** We have

$$\langle \psi, \mu_D \rangle = \langle \psi, \mu_C \rangle \cdot \left\langle \frac{\psi}{\langle \psi, \mu_C \rangle}, \mu_D \right\rangle = \langle \psi, \mu_C \rangle \cdot \left\langle \psi|_C, \mu_D|_C \right\rangle$$ ◀

where the second equality follows from $D \subseteq C$.

▶ **Lemma 8.** *Let $\mathrm{codim}(C) = k$. We have*

$$\begin{aligned}\textsc{Skew}_\psi(C) &= 2^k \Pr_{\mathbf{x}\sim\psi}[\mathbf{x} \in C] - 1 \\ &= \frac{1}{\Pr_{x\sim\mu}[\mathbf{x} \in C]} \left( \Pr_{x\sim\psi}[\mathbf{x} \in C] - \Pr_{x\sim\mu}[\mathbf{x} \in C] \right).\end{aligned}$$

**Proof.** We have

$$\begin{aligned}\langle \psi, \mu_C \rangle &= \mathop{\mathbb{E}}_{\mathbf{x}\in\mu_C}[\psi(\mathbf{x})] \\ &= \sum_{x\in C} \frac{\psi(x)}{2^{n-k}} \\ &= 2^k \sum_{x\in C} \Pr_{\mathbf{x}\sim\psi}[\mathbf{x} = x] \\ &= 2^k \Pr_{\mathbf{x}\sim\psi}[\mathbf{x} \in C].\end{aligned}$$

Hence

$$\begin{aligned}\langle \psi, \mu_C \rangle - 1 &= 2^k \Pr_{\mathbf{x}\sim\psi}[\mathbf{x} \in C] - 1 \\ &= \frac{\Pr_{\mathbf{x}\sim\psi}[\mathbf{x} \in C] - 2^{-k}}{2^{-k}}.\end{aligned}$$

Since $\Pr_{\mathbf{x}\sim\mu}[\mathbf{x} \in C] = 2^{-k}$, the claim follows. ◀

▶ **Lemma 10.** *For any $K \subseteq [n]$, we have*

$$\sum_{\substack{D=(K,w)\\ w\in\{\pm1\}^K}} \textsc{Skew}(D) = 0.$$

**Proof.** Consider the sum:

$$\sum_{\substack{D=(K,w)\\ w\in\{\pm1\}^K}} \textsc{Skew}(D) = \sum_{\substack{D=(K,w)\\ w\in\{\pm1\}^K}} 2^k \left( \Pr_{x\sim\psi}[\mathbf{x}\in D] - \Pr_{x\sim\mu}[\mathbf{x}\in D] \right) = 0.$$

The first equality comes from Lemma 8, the second follows since the set of cubes $D$ form a partition of $\{\pm1\}^n$. ◀

▶ **Lemma 11.** *If $\{C_1,\ldots,C_{2^\ell}\}$ is a partition of $C$, then*

$$\textsc{Skew}_\psi(C) = \frac{1}{2^\ell}\sum_{i=1}^{2^\ell} \textsc{Skew}_\psi(C_i).$$

**Proof.** We have

$$\begin{aligned}
\textsc{Skew}_\psi(C) &= 2^k \left( \Pr_{\mathbf{x}\sim\psi}[\mathbf{x}\in C] - \frac{1}{2^k} \right) \\
&= \frac{2^{k+\ell}}{2^\ell} \left( \sum_{i=1}^{2^\ell} \left( \Pr_{\mathbf{x}\sim\psi}[\mathbf{x}\in C_i] - \frac{1}{2^{k+\ell}} \right) \right) \\
&= \frac{1}{2^\ell} \left( \sum_{i=1}^{2^\ell} 2^{k+\ell} \left( \Pr_{\mathbf{x}\sim\psi}[\mathbf{x}\in C_i] - \frac{1}{2^{k+\ell}} \right) \right) \\
&= \frac{1}{2^\ell}\sum_{i=1}^{2^\ell} \textsc{Skew}_\psi(C_i).
\end{aligned}$$

◀

▶ **Corollary 18.** *For $J\subseteq[n]$ and a distribution $\psi$ over $\{\pm1\}^n$,*

$$W^{\le k}(\psi, J) \le 2^{|J|} e^2 (\ln(e\,\|\psi\|_\infty))^k.$$

**Proof.** Recall that

$$W^{\le k}(\psi, J) = \sum_{T\subseteq J} \sum_{\substack{S\subseteq[n]\setminus J\\ |S|\le k}} \widehat{\psi}(S\cup T)^2$$

We will show that for any $T\subseteq J$, we have

$$\sum_{\substack{S\subseteq[n]\setminus J\\ |S|\le k}} \widehat{\psi}(S\cup T)^2 \le e^2(\ln(e\,\|\psi\|_\infty))^k. \tag{8}$$

The claim will then follow by summing over all $2^{|J|}$ choices of $T\subseteq J$. To prove Equation (8), we define $\psi^{(T)}:\{\pm1\}^{[n]\setminus J}\to\mathbb{R}$ as

$$\psi^{(T)}(x) = \mathop{\mathbb{E}}_{\mathbf{z}\sim\{\pm1\}^J}[\psi(x\circ\mathbf{z})\chi_T(\mathbf{z})].$$

Note that unlike $\psi$, $\psi^{(T)}$ can be negative. By orthogonality of characters, we have

$$\psi^{(T)}(x) = \sum_{S \subseteq [n] \setminus J} \widehat{\psi}(S \cup T)\chi_S(x).$$

Since $\psi^{(T)}$ is a signed average of $\psi$ which is non-negative, we have $\left\|\psi^T\right\|_\infty \leq \|\psi\|_\infty$ and

$$\begin{aligned}\left\|\psi^T\right\|_1 &= \mathop{\mathbb{E}}_{\mathbf{x} \in \{\pm 1\}^{[n]\setminus J}} \left[\left| \mathop{\mathbb{E}}_{\mathbf{z} \sim \{\pm 1\}^J} [\psi(x \circ \mathbf{z})\chi_T(\mathbf{z})] \right|\right] \\ &\leq \mathop{\mathbb{E}}_{\mathbf{x} \in \{\pm 1\}^{[n]\setminus J}} \left[\left| \mathop{\mathbb{E}}_{\mathbf{z} \sim \{\pm 1\}^J} [|\psi(x \circ \mathbf{z})|] \right|\right] = \mathop{\mathbb{E}}_{\mathbf{x} \in \{\pm 1\}^n} [|\psi(\mathbf{x})|] = 1.\end{aligned}$$

The proof of Theorem 17 only uses bounds on the 1 and $\infty$ norms of $\psi$. Hence we can repeat the same proof with $\psi^{(T)}$ to get an identical bound. This proves Equation (8). ◀

▶ **Lemma 19.** *For $P \subseteq [n]$ and a distribution $\psi$ over $\{\pm 1\}^n$, the restriction $\psi_P$ is given by*

$$\psi_P(y) = \sum_{S \subseteq P} \widehat{\psi}(S)\chi_S(y).$$

**Proof.** Let $|P| = p < n$. For $x \in \{\pm 1\}^P$, we have

$$\psi_P(y) = 2^p \Pr_{\mathbf{x} \sim \psi}[\mathbf{x}_P = y] = \sum_{z \in \{\pm 1\}^{\bar{P}}} \frac{\psi(y \circ z)}{2^{n-p}} = \mathop{\mathbb{E}}_{\mathbf{z} \sim \mu_{\bar{P}}} [\psi(y \circ \mathbf{z})].$$

where the last expectation is over the bits $\mathbf{z}$ assigned to $\bar{P}$ being chosen uniformly at random. Using the Fourier expansion of $\psi$,

$$\mathop{\mathbb{E}}_{\mathbf{z} \sim \mu_{\bar{P}}} [\psi(y \circ \mathbf{z})] = \sum_{S \subseteq [n]} \widehat{\psi}(S) \mathop{\mathbb{E}}_{\mathbf{z} \sim \mu_{\bar{P}}} \chi_S(y \circ \mathbf{z}) = \sum_{S \subseteq P} \widehat{\psi}(S)\chi_S(y).$$ ◀

▶ **Lemma 20.** *Let $P \subsetneq [n]$. Let $\psi'$ be a distribution on $\{\pm 1\}^P$. Define a distribution $\psi$ on $\{\pm 1\}^n$ by $\psi(x) = \psi'(x_P)$. Then*

1. *$\psi$ is the product distribution of $\psi'$ with the uniform distribution on $\{\pm 1\}^{\bar{P}}$.*
2. *$\|\psi\|_\infty = \|\psi'\|_\infty$.*
3. *$C$ is a minimal skewed subcube under $\psi$ iff it is a minimal skewed subcube under $\psi'$.*

**Proof.** It follows that $\psi$ is a distribution since it is non-negative and $\|\psi\|_1 = 1$. Since the uniform distribution on $\{\pm 1\}^{(\bar{P})}$ is given by $\mu_{\bar{P}}(y) = 1$ for all $y \in \{\pm 1\}^{\bar{P}}$, it follows that $\psi = \psi' \times \mu_{\bar{P}}$ is the product of $\psi'$ and $\mu_{\bar{P}}$.

Claim (2) follows trivially from the definition of $\psi$.

For claim (3), we show that if $C = (K, y)$ is a minimal skewed subcube under $\psi$, then $K \subseteq P$. Indeed, suppose $i \in K \setminus P$. Consider the parent subcube $D \supsetneq C$ obtained by *freeing* the coordinate $i$. Then $\Pr_{\mathbf{x} \in \psi}[\mathbf{x} \in C] = \Pr_{\mathbf{x} \in \psi}[\mathbf{x} \in D]/2$, whereas $\operatorname{codim}(C) = \operatorname{codim}(D) + 1$, hence

$$\textsc{Skew}_\psi(C) = 2^{\operatorname{codim}(C)} \Pr_{\mathbf{x} \in \psi}[\mathbf{x} \in C] - 1 = \textsc{Skew}_\psi(D).$$

This violates the definition of minimality. In the other direction, it is easy to see that a minimal skewed subcube under $\psi'$ is also a minimal skewed subcube under $\psi$. ◀

▶ **Lemma 21.** *Let $C = (J, z)$. Then*

$$\psi|_C(x) = \sum_{S \subseteq [n] \setminus J} \chi_S(x) \frac{\sum_{T \subseteq J} \widehat{\psi}(S \cup T)\chi_T(z)}{\langle \psi, \mu_C \rangle}.$$

**Proof.** Recall that $\psi|_C(x) = \psi(x)/\langle \psi, \mu_C \rangle$ for $x \in C$. For $x \in \{\pm 1\}^{[n]\setminus J}$ let $x \circ z$ denote the string obtained by setting coordinates in $J$ to $z$ and those in $[n] \setminus J$ to $x$. Then

$$\begin{aligned}
\psi|_C(x) &= \frac{\psi(x \circ z)}{\langle \psi, \mu_C \rangle} \\
&= \frac{1}{\langle \psi, \mu_C \rangle} \sum_{S \subset [n]\setminus J} \sum_{T \subset J} \chi_{S \cup T}(x \circ z)\widehat{\psi}(S \cup T) \\
&= \frac{1}{\langle \psi, \mu_C \rangle} \sum_{S \subset [n]\setminus J} \sum_{T \subset J} \chi_S(x)\chi_T(z)\widehat{\psi}(S \cup T) \\
&= \sum_{S \subset [n]\setminus J} \chi_S(x) \frac{\sum_{T \subset J} \chi_T(z)\widehat{\psi}(S \cup T)}{\langle \psi, \mu_C \rangle}
\end{aligned}$$

◀

▶ **Lemma 32** ([19]). *Let $d \geq 2$ be even and let $n + 1 = 2^l \geq d$. There exists a $\mathbb{F}_2$-linear code $\mathcal{C}_{BCH} \subseteq \{0,1\}^n$ with minimum distance $d$, which contains $\Omega(n^{d/2+1})$ minimum weight codewords.*

**Proof.** Let $d = 2e + 2$ for $e \geq 0$. We use the fact that BCH codes are $[n, n - el - 1, 2e + 2)]$ codes [2, Theorem 16.21]. We need to show that there are many minimum weight codewords. For this, let us consider the parity check matrix $H$ which has dimension $(el + 1) \times n$ so that $\mathcal{C}_{BCH} = \{x \in \mathbb{F}_2^n : Hx = 0\}$.

Now let us consider the mapping $x \to Hx$ for all $x : \mathrm{wt}(x) = e + 1$. This maps each $x$ to a vector $y \in \{0,1\}^{el+1}$. For each $y \in \{0,1\}^{el+1}$, let $b_y = |\{x : \mathrm{wt}(x) = e + 1, Hx = 0\}|$ be the number of vectors of weight $e + 1$ mapped to $y$. Then we have $\sum_y b_y = \binom{n}{e+1}$, hence

$$\sum_y b_y^2 \geq \frac{\left(\sum_y b_y\right)^2}{2^{el+1}} = \frac{\binom{n}{e+1}^2}{2(n+1)^e}.$$

For $x_1 \neq x_2$ both of weight $e + 1$, if $Hx_1 = Hx_2$ then $H(x_1 + x_2) = 0$, hence $x_1 + x_2$ is a non-zero codeword of weight at most $2e + 2$, hence it is in fact a minimum weight codeword. Since there are $\binom{2e+2}{e+1} < 2^{2e+2}$ ways to write each vector of weight $2e + 2$ as such as sum, hence the number of vectors of codewords of weight $2e + 2$ is at least

$$\frac{1}{2^{2e+2}} \sum_y \binom{b_y}{2} = \frac{1}{2^{2e+3}} \sum_y (b_y^2 - b_y) \geq \frac{1}{2^{2e+3}} \left( \frac{\binom{n}{e+1}^2}{2(n+1)^e} - \binom{n}{e+1} \right) = \Omega_e(n^{e+2}).$$ ◀

▶ **Theorem 34** (Algorithm for Positive Skew). *Given sample access to a distribution $\psi$ on $\{\pm 1\}^n$, integer $k \leq n$, and parameters $\gamma \in (0, 2^k - 1]$, $\epsilon \in (0, 1]$ and $\lambda \in [0, 1]$, there is an algorithm that returns all $(\gamma, \epsilon)$-minimal skewed subcubes of codimension at most $k$ in time:*

$$\tilde{O}\left(n^{k\left(\frac{\omega}{3-\lambda}\right)}\right) + k^{O(k)} \cdot (\ln(e \|\psi\|_\infty))^k \left( \frac{\tilde{O}(n^{k/3})}{(\epsilon\sqrt{\gamma})^{4/\lambda}} + \frac{\mathrm{poly}(n)}{(\epsilon\sqrt{\gamma})^{2k}} \right)$$

*where $\omega$ is the matrix multiplication exponent, and $\tilde{O}$ hides* $\mathrm{poly}\log n$ *factors.*

▶ **Theorem 35** (Algorithm for Negative Skew). *Given sample access to a distribution $\psi$ on $\{\pm 1\}^n$, integer $k \leq n$, and parameters $\gamma \in (0,1]$, $\epsilon \in (0,1]$ and $\lambda \in [0,1]$, there is an algorithm that returns all $(-\gamma,\epsilon)$-minimal skewed subcubes of codimension at most $k$ in time:*

$$\tilde{O}\left(n^{k\left(\frac{\omega}{3-\lambda}\right)}\right) + k^{O(k)} \cdot (\ln(e\,\|\psi\|_\infty))^k \left(\frac{\tilde{O}(n^{k/3})}{(\epsilon\gamma)^{4/\lambda}} + \frac{\mathrm{poly}(n)}{(\epsilon\gamma)^{2k}}\right)$$

*where $\omega$ is the matrix multiplication exponent, and $\tilde{O}$ hides* $\mathrm{poly}\log n$ *factors.*

The algorithm is FindSkew$^+(\emptyset,\emptyset)$ in the positive case and FindSkew$^-(\emptyset,\emptyset)$ in the negative case with the nondeterminism replaced. In the positive case, one could replace the enumeration of Fourier coefficients (this is Line 6 in Algorithm 1 and Line 4 in Algorithm 2) by a call to FindFourierCoefficients. However this would naively yield a running time bound of $O(n^{0.8k}) \cdot k^{O(k)} \left(\ln(e\,\|\psi\|_\infty/\epsilon\gamma)\right)^k/(\epsilon\gamma)^{2k+O(1)}$. We show the stronger bound claimed in the theorem by making the following modification. Instead of running FindFourierCoefficients at every recursive call, we run it once at the top level to get the list $L$ of heavy coefficients for $\psi$, and "deduce" the heavy Fourier coefficients for each restricted distribution $\psi|_C$ from the original list $L$. To do so efficiently, we require a data structure which we now explain.

We preprocess $L$ by creating a graph $G_L$. Vertices of this graph are indexed by elements of the power set $2^{[n]}$. For each coefficient $S \in L$, and each subset $T \subset S$, add the directed edge $T \to S$. Furthermore, each $T$ stores $k$ lists, where the $i^{th}$ list contains all sets $S$ in the out-neighborhood of $T$ such that $|S \backslash T| = i$. Since $2^k|L|$ is a bound on both the total number of edges and the total number of vertices in the graph, creating the graph takes $O(2^k|L|)$ time. Creating the partitions of the out-neighborhoods also takes $O(2^k|L|)$ times since each edge in the graph need only be processed once.

We summarize the algorithm below for reference.

**Algorithm 4** PreprocessCoefficients$(L)$.

```
V, E ← ∅
for S ∈ L do
    for T ⊆ S do
        V ← V ∪ {S, T}
        E ← E ∪ {(T → S)}
for edge (T → S) ∈ E do
    Let i = |S\T|, and add S to the i^th list stored at T.
return G_L = (V, E)
```

Next, given $G_L$ the preprocessed form of $L$, a target subcube $C = (J, z)$ and a threshold $\tau$, we show how to output all Fourier coefficients $\widehat{\psi|_C}(S)$ such that $\widehat{\psi|_C}(S) \geq \tau$.

▶ **Lemma 41.** *Let $L = \{S \subseteq [n] : |S| \leq k, |\widehat{\psi}(S)| \geq \tau/4^k\}$, and $G_L$ be the output of* Preprocess$(L)$. *Then* DeduceSubcubeCoefficients$(G_L, C, \tau)$ *returns the list of Fourier coefficients of $\psi|_C$ of degree at most $k$ and magnitude at least $\tau$. Furthermore,* DeduceSubcubeCoefficients$(G_L, C, \tau)$ *runs in time*

$$\mathrm{poly}(n) \cdot O(|L'|) \leq \mathrm{poly}(n) \cdot 2^{O(k)} \cdot \frac{(\ln(e\,\|\psi\|_\infty))^{k-|J|}}{\tau^2}.$$

**Algorithm 5** DeduceSubcubeCoefficients$(G_L, C, \tau)$.

1: Let $J$ be the coordinates fixed by $C$.
2: $L' \leftarrow \emptyset$
3: **for** $T \subseteq J$ **do**
4: | **for** $S$ out-neighbor of $T$ in $G_L$ with $|S \backslash T| \leq k - |J|$ **do**
5: | | Check by sampling if $\left|\widehat{\psi|_C}(S \backslash J)\right| \geq 3\tau/4$ to accuracy $\tau/4$. If so, add $S \backslash J$ to $L'$.
6: **return** $L'$.

**Proof.** The output $L'$ consists only of sets $S'$ such that $S' \cup J \in L$, $S' \cap J = \emptyset$ and $|S'| \leq k - |J|$. Furthermore it contains all $S'$ meeting these criteria such that $\left|\widehat{\psi|_C}(S')\right| \geq \tau$, note that for any set $H \in L$ such that $|H| \leq k$, if $T$ is chosen to be $H \cap J$, then $|H \backslash J| = |H \backslash T| \leq k - |J|$ and thus the algorithm will add $H \backslash J$ to the output.

To obtain the claim of the lemma, we need to argue that for every Fourier coefficient $\widehat{\psi|_C}(S)$ of absolute value at least $\tau$, the set $S$ must appear in $L'$. We need the following consequence of Lemma 21: let $\psi$ be a distribution on $\{\pm 1\}^n$ and let $C = (J, z)$ be a subcube. Then

$$\widehat{\psi|_C}(S) = \sum_{T \subseteq J} \frac{\chi_T(z)\widehat{\psi}(S \cup T)}{\langle \psi, \mu_C \rangle}.$$

It follows that every coefficient $\widehat{\psi|_C}(S)$ is the signed sum of at most $2^k$ coefficients $\widehat{\psi}(R) = \widehat{\psi}(S \cup T)$, which is then scaled by $1/\langle \psi, \mu_C \rangle$ (at most $2^k$). Furthermore, if $|S| \leq k - |J|$, then each such coefficient has $|R| \leq |S| + |T| \leq k$. Thus if $\widehat{\psi|_C}(S) \geq \tau$, there is at least one $R \subseteq J$ with $|R| \leq k$ such that $|\widehat{\psi}(R)| \geq \tau/4^k$, which means that $R \in L$.

Finally, the running time claim follows from Corollary 18 and the fact that for every $S \in L$ we have $\widehat{\psi}(S) \geq \tau/4^k$. ◀

We are now ready to prove our main theorem. We start with the positive case.

**Proof of Theorem 34.** We start by running FindFourierCoefficients$(\psi, k, \rho^+, \lambda)$ once, with $\rho^+ = \epsilon\sqrt{\gamma}/16^k$. This outputs a list $L^+$ containing all $S$ where $|S| \leq k$ and $\widehat{\psi}(S) \geq \epsilon\sqrt{\gamma}/4^k$. Subsequently, we compute $G_{L^+} \leftarrow$ PreprocessCoefficients$(L^+)$ using the output. This set up phase has running time $R^+$ bounded by

$$R^+ \leq \tilde{O}\left(n^{k\left(\frac{\omega}{3-\lambda}\right)}\right) + \frac{k^{O(k)}\tilde{O}\left(n^{k/3}\right)\ln(e\,\|\psi\|_\infty)^k}{(\epsilon\sqrt{\gamma})^{4/\lambda}}.$$

Next we run FindSkew$^+(\emptyset, \emptyset)$ but we replace the nondeterministic enumeration of Fourier coefficients (this is Line 6 in Algorithm 1) by DeduceSubcubeCoefficients$(G_{L^+}, C, \tau^+)$ where $\tau^+ := \epsilon\sqrt{\gamma}/\left(k_t \cdot \binom{k_t}{|S_t|}\right)$. We also replace the nondeterministic choice of $z$ (Line 7 in Algorithm 1) by simple enumeration over all possible choices. Correctness follows from Lemma 24 together with Lemmas 39 and 41 and it remains to show the running time bound.

By Lemma 41, at every subcube $C$ the algorithm spends time at most

$$\mathrm{poly}(n) \cdot 2^{O(k)} \cdot \frac{(\ln(e\,\|\psi\|_\infty))^{k-|J|}}{(\tau^+)^2} = \mathrm{poly}(n) \cdot 2^{O(k)} \cdot \frac{(\ln(e\,\|\psi\|_\infty))^{k-|J|}}{(\epsilon\sqrt{\gamma})^2}$$

to run DeduceSubcubeCoefficients.

On the other hand, the proof of Lemma 26 requires that the branching factor of FindSkew$^+$ be bounded as in Lemma 25. Since this bound is at least $(\ln(e\,\|\psi\|_\infty))^{k-|J|}/(\epsilon\sqrt{\gamma})$ (and we may assume WLOG that the branching factor is *at least* this threshold), we may amortize the cost of each call to DeduceSubcubeCoefficients by charging to each child call the average running time per child. The time spent per child is $\mathrm{poly}(n)\cdot 2^{O(k)}$, and we argued in Lemma 26 that the total number of recursive calls to FindSkew$^+$ is at most

$$k^{O(k)}\left(\frac{\ln(e\,\|\psi\|_\infty)}{\epsilon^2\gamma}\right)^k.$$

Thus we may bound the running time of FindSkew$^+(\emptyset,\emptyset)$ by this expression as well.

To conclude, the final running time of the algorithm in the positive skew case is:

$$\begin{aligned}
&R^+ + \mathrm{poly}(n,k^k)\cdot k^{O(k)}\left(\frac{\ln(e\,\|\psi\|_\infty)}{\epsilon^2\gamma}\right)^k\\
&\le \tilde{O}\left(n^{k\left(\frac{\omega}{3-\lambda}\right)}\right) + \frac{k^{O(k)}\tilde{O}\left(n^{k/3}\right)\ln(e\,\|\psi\|_\infty)^k}{(\epsilon\sqrt{\gamma})^{4/\lambda}} + \mathrm{poly}(n,k^k)\cdot k^{O(k)}\left(\frac{\ln(e\,\|\psi\|_\infty)}{\epsilon^2\gamma}\right)^k\\
&\le \tilde{O}\left(n^{k\left(\frac{\omega}{3-\lambda}\right)}\right) + k^{O(k)}\cdot(\ln(e\,\|\psi\|_\infty))^k\left(\frac{\tilde{O}\left(n^{k/3}\right)}{(\epsilon\sqrt{\gamma})^{4/\lambda}} + \frac{\mathrm{poly}(n)}{(\epsilon\sqrt{\gamma})^{2k}}\right).
\end{aligned}$$

◀

The negative case is identical, but with a different setting of parameters.

**Proof of Theorem 35.** In this case we run FindFourierCoefficients$(\psi,k,\rho^-,\lambda)$ with $\rho^- = \epsilon\gamma/16^k$. This outputs the list $L^-$, and we set $G_{L^-} \leftarrow$ PreprocessCoefficients$(L^-)$. This all has running time $R^-$ bounded by

$$R^- \le \tilde{O}\left(n^{k\left(\frac{\omega}{3-\lambda}\right)}\right) + \frac{k^{O(k)}\tilde{O}\left(n^{k/3}\right)\ln(e\,\|\psi\|_\infty)^k}{(\epsilon\gamma)^{4/\lambda}}.$$

Now we run FindSkew$^-(\emptyset,\emptyset)$, with Fourier coefficient enumeration (Algorithm 2 in Algorithm 2) replaced with DeduceSubcubeCoefficients$(G_{L^-},C,\tau^-)$, and we set $\tau^- = \epsilon\gamma/\left(k_t\cdot\binom{k_t}{|S_t|}\right)$. The analysis is identical to the positive case, and the final running time is:

$$\tilde{O}\left(n^{k\left(\frac{\omega}{3-\lambda}\right)}\right) + k^{O(k)}\cdot(\ln(e\,\|\psi\|_\infty))^k\left(\frac{\tilde{O}\left(n^{k/3}\right)}{(\epsilon\gamma)^{4/\lambda}} + \frac{\mathrm{poly}(n)}{(\epsilon\gamma)^{2k}}\right).$$

◀