Linear Sketching over $\mathbb{F}_2$ *

Sampath Kannan[1]
University of Pennsylvania
kannan@cis.upenn.edu

Elchanan Mossel[2]
Massachusetts Institute of Technology
elmos@mit.edu

Swagato Sanyal[3]
Division of Mathematical Sciences, Nanyang Technological University, Singapore and Centre for Quantum Technologies, National University of Singapore, Singapore
sanyalswagato@gmail.com

Grigory Yaroslavtsev[4]
Indiana University, Bloomington
grigory@grigory.us

Abstract

We initiate a systematic study of linear sketching over $\mathbb{F}_2$. For a given Boolean function treated as $f: \mathbb{F}_2^n \to \mathbb{F}_2$ a randomized $\mathbb{F}_2$-sketch is a distribution $\mathcal{M}$ over $d \times n$ matrices with elements over $\mathbb{F}_2$ such that $\mathcal{M}x$ suffices for computing $f(x)$ with high probability. Such sketches for $d \ll n$ can be used to design small-space distributed and streaming algorithms.

Motivated by these applications we study a connection between $\mathbb{F}_2$-sketching and a two-player one-way communication game for the corresponding XOR-function. We conjecture that $\mathbb{F}_2$-sketching is optimal for this communication game. Our results confirm this conjecture for multiple important classes of functions: 1) low-degree $\mathbb{F}_2$-polynomials, 2) functions with sparse Fourier spectrum, 3) most symmetric functions, 4) recursive majority function. These results rely on a new structural theorem that shows that $\mathbb{F}_2$-sketching is optimal (up to constant factors) for uniformly distributed inputs.

Furthermore, we show that (non-uniform) streaming algorithms that have to process random updates over $\mathbb{F}_2$ can be constructed as $\mathbb{F}_2$-sketches for the uniform distribution. In contrast with the previous work of Li, Nguyen and Woodruff (STOC'14) who show an analogous result for linear sketches over integers in the adversarial setting our result does not require the stream length to be triply exponential in n and holds for streams of length $\tilde{O}(n)$ constructed through uniformly random updates.

2012 ACM Subject Classification Theory of computation → Probabilistic computation, Theory of computation → Streaming models, Theory of Computation → Computational complexity and cryptography → Communication complexity.

* This is an improved version of Kannan, Mossel, Yaroslavtsev `https://arxiv.org/pdf/1611.01879.pdf` [24] giving tight dependence on error in Theorem 4, a new result for degree-d $\mathbb{F}_2$ polynomials and including several other changes. ECCC preprint is availabe at `https://eccc.weizmann.ac.il/report/2018/064/`.

[1] This work was supported by grants NSF CICI 1547360 and ONR N00014-15-1-2006.

[2] E.M. acknowledges the support of grant N00014-16-1-2227 from Office of Naval Research and of NSF awards CCF 1320105 and DMS-1737944 as well as support from Simons Think Tank on Geometry & Algorithms.

[3] S.S. acknowledges the support by the Singapore National Research Foundation under NRF RF Award No. NRF-NRFF2013-13. Part of this work was done when S. S. was a visiting research fellow at the Tata Institute of Fundamental Research, Mumbai.

[4] This work was supported by NSF award 1657477.

33rd Computational Complexity Conference (CCC 2018).
Editor: Rocco A. Servedio; Article No. 8; pp. 8:1–8:37
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Keywords and phrases Linear sketch, Streaming algorithms, XOR-functions, Communication complexity.

Digital Object Identifier 10.4230/LIPIcs.CCC.2018.8

1 Introduction

Linear sketching is the underlying technique behind many of the biggest algorithmic breakthroughs of the past two decades. It has played a key role in the development of streaming algorithms since [3]and most recently has been the key to modern randomized algorithms for numerical linear algebra (see survey [52]), graph compression (see survey [38]), dimensionality reduction, etc. Linear sketching is robust to the choice of a computational model and can be applied in settings as seemingly diverse as streaming, MapReduce as well as various other distributed models of computation including the congested clique model [19, 12, 23], allowing to save computational time, space and reduce communication in distributed settings. This remarkable versatility is based on properties of linear sketches enabled by linearity: simple and fast updates and mergeability of sketches computed on distributed data. Compatibility with fast numerical linear algebra packages makes linear sketching particularly attractive for applications.

Even more surprisingly linear sketching over the reals is known to be the best possible algorithmic approach (unconditionally) in certain settings. Most notably, under some mild conditions linear sketches are known to be almost space optimal for processing dynamic data streams [10, 32, 1]. Optimal bounds for streaming algorithms for a variety of computational problems can be derived through this connection by analyzing linear sketches rather than general algorithms. Examples include approximate matchings [5, 4], additive norm approximation [1] and frequency moments [32, 51].

In this paper we study the power of linear sketching over $\mathbb{F}_2$. [5] To the best of our knowledge no such systematic study currently exists as prior work focuses on sketching over the field of reals (or large finite fields as reals are represented as word-size bounded integers). Formally, for a random set $\mathbf{S} \subseteq [n]$ let $\chi_{\mathbf{S}} = \bigoplus_{i \in \mathbf{S}} x_i$. Given a function $f \colon \mathbb{F}_2^n \to \mathbb{F}_2$ that needs to be evaluated over an input $x = (x_1, \ldots, x_n)$ we are looking for a distribution over k subsets $\mathbf{S}_1, \ldots, \mathbf{S}_k \subseteq [n]$ such that the following holds: for any input x given parities computed over these sets and denoted as $\chi_{\mathbf{S}_1}(x), \chi_{\mathbf{S}_2}(x), \ldots, \chi_{\mathbf{S}_k}(x)$, it should be possible to compute $f(x)$ with probability $1 - \delta$. While the switch from reals to $\mathbb{F}_2$ might seem restrictive, we are unaware of any problem for which sketching over reals gives any advantage over $\mathbb{F}_2$. Furthermore, as shown very recently and subsequently to the early version of this work [24], almost all dynamic graph streaming algorithms[6] can be seen as $\mathbb{F}_2$-sketches [26] without losing optimality in space[7].

In matrix form $\mathbb{F}_2$-sketching corresponds to multiplication over $\mathbb{F}_2$ of the row vector $x \in \mathbb{F}_2^n$ by a random $n \times k$ matrix whose i-th column is a characteristic vector of the random parity $\chi_{\mathbf{S}_i}$:

[5] It is easy to see that sketching over finite fields can be significantly better than linear sketching over integers for certain computations. As an example, consider a function ($x \mod 2$) (for an integer input x) which can be trivially sketched with 1 bit over the field of two elements while any linear sketch over the integers requires word-size memory.

[6] With the only exception being the work of [25] on spectral graph sparsification.

[7] Technically [26] uses $\mathbb{F}_3$, but replacing $\mathbb{F}_3$ with $\mathbb{F}_2$ doesn't change their results.

$$\begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} \vdots & \vdots & \vdots & \vdots \\ \chi_{\mathbf{S}_1} & \chi_{\mathbf{S}_2} & \cdots & \chi_{\mathbf{S}_k} \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} = \begin{pmatrix} \chi_{\mathbf{S}_1}(x) & \chi_{\mathbf{S}_2}(x) & \dots & \chi_{\mathbf{S}_k}(x) \end{pmatrix}$$

This sketch alone should then be sufficient for computing f with high probability for any input x. This motivates us to define the *randomized linear sketch* complexity of a function f over $\mathbb{F}_2$ as the smallest k which allows one to satisfy the above guarantee.

▶ **Definition 1** ($\mathbb{F}_2$-sketching). For a function $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$ we define its *randomized linear sketch complexity*[8] over $\mathbb{F}_2$ with error δ (denoted as $R_\delta^{lin}(f)$) as the smallest integer k such that there exists a distribution $\chi_{\mathbf{S}_1}, \chi_{\mathbf{S}_2}, \dots, \chi_{\mathbf{S}_k}$ over k linear functions over $\mathbb{F}_2$ and a postprocessing function $g : \mathbb{F}_2^k \to \mathbb{F}_2$[9] which satisfies:

$$\forall x \in \mathbb{F}_2^n\colon \Pr_{\mathbf{S}_1,\dots,\mathbf{S}_k}[f(x_1, x_2, \dots, x_n) = g(\chi_{\mathbf{S}_1}(x), \chi_{\mathbf{S}_2}(x), \dots, \chi_{\mathbf{S}_k}(x))] \geq 1 - \delta.$$

We note that while the above definition requires that f is computed exactly, most of our structural results including Theorem 4 can be extended to allow approximate computation of real-valued functions $f\colon \mathbb{F}_2^n \to \mathbb{R}$ as shown in [54].

As we show in this paper the study of $R_\delta^{lin}(f)$ is closely related to a certain communication problem. For $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$ define the XOR-function $f^+\colon \mathbb{F}_2^n \times \mathbb{F}_2^n \to \mathbb{F}_2$ as $f^+(x, y) = f(x + y)$ where $x, y \in \mathbb{F}_2^n$. Consider a communication game between two players Alice and Bob holding inputs x and y respectively. Given access to a shared source of random bits Alice has to send a single message to Bob so that he can compute $f^+(x, y)$. This is known as the one-way communication problem for XOR-functions.

▶ **Definition 2** (Randomized one-way communication complexity of XOR function). For a function $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$ the *randomized one-way communication complexity* with error δ (denoted as $R_\delta^{\to}(f^+)$) of its XOR-function is defined as the smallest size[10] (in bits) of the (randomized using public randomness) message $M(x)$ from Alice to Bob which allows Bob to evaluate $f^+(x, y)$ for any $x, y \in \mathbb{F}_2^n$ with error probability at most δ.

Communication complexity of XOR-functions has been recently studied extensively in the context of the log-rank conjecture (see e.g. [45, 55, 39, 29, 31, 47, 33, 49, 35, 18]). However, such studies either mostly focus on deterministic communication complexity or are specific to the two-way communication model. We discuss implications of this line of work for our $\mathbb{F}_2$-sketching model in our discussion of prior work.

It is easy to see that $R_\delta^{\to}(f^+) \leq R_\delta^{lin}(f)$ as using shared randomness for sampling $\mathbf{S}_1, \dots, \mathbf{S}_k$ Alice can just send k bits $\chi_{\mathbf{S}_1}(x), \chi_{\mathbf{S}_2}(x), \dots, \chi_{\mathbf{S}_k}(x)$ to Bob who can for each

[8] In the language of decision trees this can be interpreted as randomized non-adaptive parity decision tree complexity. We are unaware of any systematic study of this quantity either. Since heavy decision tree terminology seems excessive for our applications (in particular, sketching is done in one shot so there isn't a decision tree involved) we prefer to use a shorter and more descriptive name.

[9] Technically g can also depend on the sampled sets $\mathbf{S}_1, \dots, \mathbf{S}_k$, but all sketches used in this paper are oblivious to the choice of these sets.

[10] Formally the minimum here is taken over all possible protocols where for each protocol the size of the message $M(x)$ refers to the largest size (in bits) of such message taken over all inputs $x \in \mathbb{F}_2^n$. See [28] for a formal definition.

$i \in [k]$ compute $\chi_{\mathbf{S}_i}(x+y) = \chi_{\mathbf{S}_i}(x) + \chi_{\mathbf{S}_i}(y)$. This gives Bob an $\mathbb{F}_2$-sketch of f on $x+y$ and hence suffices for computing $f^+(x,y)$ with probability $1-\delta$. The main open question raised in our work is whether the reverse inequality holds (at least approximately), thus implying the equivalence of the two notions.

▶ Conjecture 3. Is it true that $R^{\rightarrow}_\delta(f^+) = \tilde{\Theta}\left(R^{lin}_\delta(f)\right)$ for every $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$ and $0 < \delta < 1/2$?

In fact all known one-way protocols for XOR-functions can be seen as $\mathbb{F}_2$-sketches so it is natural to ask whether this is always true. In this paper we further motivate this conjecture through a number of examples of classes of functions for which it holds. One important such example from the previous work is a function $Ham_{\geq k}$ which evaluates to 1 if and only if the Hamming weight of the input string is at least k. The corresponding XOR-function $Ham^+_{\geq k}$ can be seen to have one-way communication complexity of $\Theta(k \log k)$ via the small set disjointness lower bound of [9] and a basic upper bound based on random parities [20]. Conjecture 3 would imply that in order to prove a one-way disjointness lower bound it suffices to only consider $\mathbb{F}_2$-sketches.

A deterministic analog of Definition 1 requires that $f(x) = g(\chi_{\alpha_1}(x), \chi_{\alpha_2}(x), \dots, \chi_{\alpha_k}(x))$ for a fixed choice of $\alpha_1, \dots, \alpha_k \in \mathbb{F}_2^n$. The smallest value of k which satisfies this definition is known to be equal to the Fourier dimension of f denoted as $dim(f)$. It corresponds to the smallest dimension of a linear subspace of $\mathbb{F}_2^n$ that contains the entire spectrum of f (see Section 2.2 for a formal definition). In order to keep the notation uniform we also denote it as $D^{lin}(f)$. Most importantly, as shown in [39] an analog of Conjecture 3 holds without any loss in the deterministic case, i.e. $D^{\rightarrow}(f^+) = dim(f) = D^{lin}(f)$, where $D^{\rightarrow}$ denotes the deterministic one-way communication complexity. This striking fact is one of the reasons why we suggest Conjecture 3 as an open problem.

Previous work and our results

In the discussion below using Yao's principle we switch to the equivalent notion of distributional complexity of the above problems denoted as $\mathcal{D}^{\rightarrow}_\delta$ and $\mathcal{D}^{lin}_\delta$ respectively. For the formal definitions we refer to the reader to Section 2.1 and a standard textbook on communication complexity [28]. Equivalence between randomized and distributional complexities allows us to restate Conjecture 3 as $\mathcal{D}^{\rightarrow}_\delta = \tilde{\Theta}(\mathcal{D}^{lin}_\delta)$.

For a fixed distribution μ over $\mathbb{F}_2^n$ we define $\mathcal{D}^{lin,\mu}_\delta(f)$ to be the smallest dimension of an $\mathbb{F}_2$-sketch that correctly outputs f with probability $1-\delta$ over μ. Similarly for a distribution μ over $(x,y) \in \mathbb{F}_2^n \times \mathbb{F}_2^n$ we denote distributional one-way communication complexity of f with error δ as $\mathcal{D}^{\rightarrow,\mu}_\delta(f^+)$ (See Section 2 for a formal definition). Our first main result is an analog of Conjecture 3 for the uniform distribution U over (x,y) that matches the statement of the conjecture up to constant factors:

▶ **Theorem 4.** *For any $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$ it holds that $\mathcal{D}^{\rightarrow,U}_{1/9}(f^+) \geq \frac{1}{6} \cdot \mathcal{D}^{lin,U}_{1/3}(f)$.*

In order to prove Theorem 4 we introduce the notion of an *approximate Fourier dimension* (Definition 13) that extends the definition of exact Fourier dimension to allow that only $1-\epsilon$ fraction of the total "energy" in f's spectrum should be contained in the linear subspace. The key ingredient in the proof is a structural theorem, Theorem 14, that characterizes both $\mathcal{D}^{lin,U}_\delta(f)$ and $\mathcal{D}^{\rightarrow,U}_\delta(f^+)$ in terms of f's approximate Fourier dimension.

Using Theorem 14 we confirm Conjecture 3 for several well-studied classes of functions in Section 4. It is important to note that while we could have stated these results for randomized one-way communication it is critical that all lower bounds in this section hold for uniform distribution in order to derive our results for random streams in Section 5.

Low-degree $\mathbb{F}_2$ polynomials

Low-degree $\mathbb{F}_2$ polynomials have been extensively studied in theoretical computer science in various contexts: learning theory (Mossel, O'Donnell and Servedio [40]), property testing (Rubinfield and Sudan [42], Bhattacharyya *et al.* [6], Alon *et al* [2]), pseudorandomness (Bogdanov and Viola [8], Lovett [34], Viola [50]), communication complexity (Tsang *et al.*[49]), etc.

Tsang *et al.* [49] studied deterministic two-way communication protocols for XOR-functions with low $\mathbb{F}_2$-degree. They gave an upper bound on deterministic communication complexity of f^+ in terms of the spectral norm and the $\mathbb{F}_2$-degree of f. Their result was obtained by observing that the communication complexity of f^+ is bounded above by the parity decision tree complexity of f, and then bounding the latter. In this work, we prove a lower bound on the randomized one-way communication complexity of f^+ in terms of the Fourier dimension of f and the $\mathbb{F}_2$-degree of f, denoted as d. We prove the following result:

$$D^{lin}(f) = O\left(R^{\rightarrow}_{1/3}(f^+) \cdot d\right).$$

In the regime $d = O(1)$, the above result implies that use of randomness does not enable us to design a better linear-sketching or a one-way communication protocol. Furthermore, since $R^{lin}_{1/3}(f) \leq D^{lin}(f)$, the above result implies Conjecture 3 for constant degree $\mathbb{F}_2$-polynomials. For $\mathbb{F}_2$ polynomials with bounded spectral norm this implies a new bound on Fourier dimension shown in Corollary 23: $D^{lin}(f) = dim(f) = O(d\|\hat{f}\|_1^2)$ improving a result of Tsang et al. for $d = \omega\left(\log^{1/3} \|\hat{f}\|_1\right)$.

Address function and Fourier sparsity

The number s of non-zero Fourier coefficients of f (known as Fourier sparsity) is one of the key quantities in the analysis of Boolean functions. It also plays an important role in the recent work on log-rank conjecture for XOR-functions [49, 46]. A recent result by Sanyal [44] shows that for Boolean functions $dim(f) = O(\sqrt{s}\log s)$, namely all non-zero Fourier coefficients are contained in a subspace of a polynomially smaller dimension. This bound is almost tight as the *address function* (see Section 4.2 for a definition) exhibits a quadratic gap. A direct implication of Sanyal's result is a deterministic $\mathbb{F}_2$-sketching upper bound of $O(\sqrt{s}\log s)$ for any f with Fourier sparsity s. As we show in Section 4.2 this dependence on sparsity can't be improved even if randomization is allowed.

Symmetric functions

A function f is symmetric if it only depends on the Hamming weight of its input. In Section 4.3 we show that Conjecture 3 holds for all symmetric functions which are not too close to a constant function or the parity function $\sum_i x_i$, where the sum is taken over $\mathbb{F}_2$.

Composition theorem for recursive majority

As an example of a composition theorem we give such a theorem for recursive majority. For an odd integer n the majority function Maj_n is defined to be 1 if and only if the Hamming weight of the input is greater than $n/2$. Of particular interest is the recursive majority function $Maj_3^{\circ k}$ that corresponds to k-fold composition of Maj_3 for $k = \log_3 n$. This function was introduced by Boppana [43] and serves as an important example of various properties of Boolean functions, most importantly in randomized decision tree complexity

([43, 22, 37, 30, 36]), deterministic parity decision tree complexity [7] and communication complexity [22, 13].

In Section 4.4 we use Theorem 14 to obtain the following result:

▶ **Theorem 5.** *For any $\epsilon \in [0, \frac{1}{2}]$, $\xi > 4\epsilon^2$ and $k = \log_3 n$ it holds that:*

$$\mathcal{D}^{\to,U}_{\frac{1-\xi}{6}}(Maj_3^{\circ k^+}) = \Omega(\epsilon^2 n).$$

Applications to streaming and distributed computing

In the turnstile streaming model of computation a vector x of dimension n is updated through a sequence of additive updates applied to its coordinates and the goal of the algorithm is to be able to output $f(x)$ at any point during the stream while using space that is sublinear in n. In the real-valued case we have either $x \in [0, m]^n$ or $x \in [-m, m]^n$ for some universal upper bound m and updates can be increments or decrements to x's coordinates of arbitrary magnitude.

For $x \in \mathbb{F}_2^n$ additive updates have a particularly simple form as they always flip the corresponding coordinate of x. In the streaming literature this model is referred to as the XOR update model (see e.g. [48]) Note that XOR updates can't be handled using standard turnstile streaming algorithms as only the coordinate but not the sign of the update is given. As we show in Section 5.2 it is easy to see based on the recent work of [10, 32, 1] that in the adversarial streaming setting the space complexity of turnstile streaming algorithms over $\mathbb{F}_2$ is determined by the $\mathbb{F}_2$-sketch complexity of the function of interest. However, this proof technique only works for very long streams which are unrealistic in practice – the length of the adversarial stream has to be triply exponential in n in order to enforce linear behavior. Large stream length requirement is inherent in the proof structure in this line of work and while one might expect to improve triply exponential dependence on n at least an exponential dependence appears necessary, which is a major limitation of this approach.

As we show in Section 5.1 it follows directly from our Theorem 4 that turnstile streaming algorithms that achieve low error probability under random $\mathbb{F}_2$ updates might as well be $\mathbb{F}_2$-sketches. For two natural choices of the random update model short streams of length either $O(n)$ or $O(n \log n)$ suffice for our reduction. We stress that our lower bounds are also stronger than the worst-case adversarial lower bounds as they hold under an average-case scenario. Furthermore, our Conjecture 3 would imply that space optimal turnstile streaming algorithms over $\mathbb{F}_2$ have to be linear sketches for adversarial streams of length only $2n$. We believe that such result will also help show an analogous statement for real-valued linear sketches thus removing the triply exponential in n stream length assumption of [32, 1].

By linearity all $\mathbb{F}_2$-sketching upper bounds are also applicable in the distributed setting where two parties Alice and Bob need to send messages to the coordinator who is required to output f^+. This is also known as the Simultaneous Message Passing (SMP) model and all our one-way lower bounds hold in this model as well.

Other previous work

Closely related to ours is work on communication protocols for XOR-functions [45, 39, 49, 18]. In particular [39] presents two basic one-way communication protocols based on random parities. The first one, stated as Fact 20 generalizes the classic communication protocol for equality. The second one uses the result of Grolmusz [17] and implies that ℓ_1-sampling of Fourier characters gives a randomized $\mathbb{F}_2$-sketch of size $O(\|\hat{f}\|_1^2)$ (for constant error).

In [18] structural results about deterministic two-way communication protocols for XOR-functions have been obtained. In particular, they show that the parity decision tree complexity of f is $O(D(f^+)^6)$. The key difference between our work and [18] lies in our focus on randomized protocols. In [18] it is left as the main open problem whether randomized parity decision tree complexity can be bounded by $poly(R(f^+))$. Our results can be seen as a step towards resolving this open problem in one-way communication setting. Full resolution of Conjecture 3 would show that the conjecture of [18] holds even without polynomial loss for one-way communication as we show for all the classes considered in Section 4.

Another line of work that is closely related to ours is the study of the two-player simultaneous message passing model (SMP). This model can also allow to prove lower bounds on $\mathbb{F}_2$-sketching complexity. Since our results hold for one-way communication they also hold in the SMP model. Moreover, in the context of our work there is no substantial difference as for product distributions the two models are essentially equivalent. Recent results in the SMP model include [39, 31, 33].

While decision tree literature is not directly relevant to us since our model doesn't allow adaptivity we remark that there has been interest recently in the study of (adaptive) deterministic parity decision trees [7] and non-adaptive deterministic parity decision trees [46, 44]. As mentioned above, our model can be interpreted as non-adaptive randomized parity decision trees and to the best of our knowledge it hasn't been studied explicitly before. Another related model is that of *parity kill numbers*. In this model a composition theorem has recently been shown by [41] but the key difference is again adaptivity.

Finally recent developements in the line of work on lifting theorems such as [15, 14] might suggest that such results might be applied in our context. However for our purposes we would need a lifting theorem for the XOR gadget and to the best of our knowledge no such result is known for randomized one-way communication.

Organization

The rest of this paper is organized as follows. In Section 2 we introduce the required background from communication complexity and Fourier analysis of Boolean functions. In Section 3 we prove Theorem 4. In Section 4 we give applications of this theorem for recursive majority (Theorem 5), address function, low-degree $\mathbb{F}_2$ polynomials and symmetric functions. In Section 5 we describe applications to streaming.

In Appendix B we give some basic results about deterministic $\mathbb{F}_2$-sketching (or Fourier dimension) of composition and convolution of functions. We also present a basic lower bound argument based on affine dispersers. In Appendix C we give some basic results about randomized $\mathbb{F}_2$-sketching including a lower bound based on extractors and a classic protocol based on random parities which we use as a building block in our sketch for LTFs. We also present evidence for why an analog of Theorem 14 doesn't hold for arbitrary distributions. In Appendix D we show a lower bound for one-bit protocols making progress towards resolving Conjecture 3.

2 Preliminaries

For an integer n we use notation $[n] = \{1, \dots, n\}$. For integers $n \le m$ we use notation $[n, m] = \{n, \dots, m\}$. For an arbitrary domain $\mathcal{D}$ we denote the uniform distribution over this domain as $U(\mathcal{D})$. We use the notation $x, x' \sim U(\mathcal{D})$ to denote that x and x' are sampled uniformly at random and independently from $\mathcal{D}$. The variance of a random variable X is denoted by $\mathsf{Var}[X]$. For a vector x and $p \geq 1$ we denote the p-norm of x as $\|x\|_p$ and reserve the notation $\|x\|_0$ for the Hamming weight.

2.1 Communication complexity

Consider a function $f\colon \mathbb{F}_2^n \times \mathbb{F}_2^n \to \mathbb{F}_2$ and a distribution μ over $\mathbb{F}_2^n \times \mathbb{F}_2^n$. The *one-way distributional complexity* of f with respect to μ, denoted as $\mathcal{D}_\delta^{\to,\mu}(f)$ is the smallest communication cost of a one-way deterministic protocol that outputs $f(x,y)$ with probability at least $1-\delta$ over the inputs (x,y) drawn from the distribution μ. The *one-way distributional complexity* of f denoted as $\mathcal{D}_\delta^{\to}(f)$ is defined as $\mathcal{D}_\delta^{\to}(f) = \sup_\mu \mathcal{D}_\delta^{\to,\mu}(f)$. By Yao's minimax theorem [53] it follows that $R_\delta^{\to}(f) = \mathcal{D}_\delta^{\to}(f)$. *One-way communication complexity over product distributions* is defined as $\mathcal{D}_\delta^{\to,\times}(f) = \sup_{\mu=\mu_x\times\mu_y} \mathcal{D}_\delta^{\to,\mu}(f)$ where μ_x and μ_y are distributions over $\mathbb{F}_2^n$.

With every two-party function $f\colon \mathbb{F}_2^n \times \mathbb{F}_2^n$ we associate a *communication matrix* $M^f \in \mathbb{F}_2^{2^n \times 2^n}$ with entries $M^f_{x,y} = f(x,y)$. We say that a deterministic protocol $M(x)$ with length t of the message that Alice sends to Bob partitions the rows of this matrix into 2^t *combinatorial rectangles* where each rectangle contains all rows of M^f corresponding to the same fixed message $y \in \{0,1\}^t$.

2.2 Fourier analysis

We consider functions[11] from $\mathbb{F}_2^n$ to $\mathbb{R}$. For any fixed $n \geq 1$, the space of these functions forms an inner product space with the inner product $\langle f, g\rangle = \mathbb{E}_{x\in\mathbb{F}_2^n}[f(x)g(x)] = \frac{1}{2^n}\sum_{x\in\mathbb{F}_2^n} f(x)g(x)$. The ℓ_2 norm of $f : \mathbb{F}_2^n \to \mathbb{R}$ is $\|f\|_2 = \sqrt{\langle f, f\rangle} = \sqrt{\mathbb{E}_x[f(x)^2]}$ and the ℓ_2 distance between two functions $f, g : \mathbb{F}_2^n \to \mathbb{R}$ is the ℓ_2 norm of the function $f - g$. In other words, $\|f-g\|_2 = \sqrt{\langle f-g, f-g\rangle} = \sqrt{\frac{1}{2^n}\sum_{x\in\mathbb{F}_2^n}(f(x)-g(x))^2}$.

For $\alpha \in \mathbb{F}_2^n$, the *character* $\chi_\alpha : \mathbb{F}_2^n \to \{+1,-1\}$ is the function defined by $\chi_\alpha(x) = (-1)^{\alpha\cdot x}$. Characters form an orthonormal basis as $\langle \chi_\alpha, \chi_\beta\rangle = \delta_{\alpha\beta}$ where δ is the Kronecker symbol. The *Fourier coefficient* of $f : \mathbb{F}_2^n \to \mathbb{R}$ corresponding to α is $\hat{f}(\alpha) = \mathbb{E}_x[f(x)\chi_\alpha(x)]$. The *Fourier transform* of f is the function $\hat{f} : \mathbb{F}_2^n \to \mathbb{R}$ that returns the value of each Fourier coefficient of f. We use notation $\mathit{Spec}(f) = \{\alpha \in \mathbb{F}_2^n : \hat{f}(\alpha) \neq 0\}$ to denote the set of all non-zero Fourier coefficients of f. The Fourier ℓ_1 norm, or the *spectral norm* of f, is defined as $\|\hat{f}\|_1 := \sum_{\alpha\in\mathbb{F}_2^n} |\hat{f}(\alpha)|$.

► Fact 6 (Parseval's identity). For any $f : \mathbb{F}_2^n \to \mathbb{R}$ it holds that

$$\|f\|_2 = \|\hat{f}\|_2 = \sqrt{\sum_{\alpha\in\mathbb{F}_2^n} \hat{f}(\alpha)^2}.$$

Moreover, if $f : \mathbb{F}_2^n \to \{+1,-1\}$ then $\|f\|_2 = \|\hat{f}\|_2 = 1$.

We use notation $A \leq \mathbb{F}_2^n$ to denote the fact that A is a linear subspace of $\mathbb{F}_2^n$.

► **Definition 7** (Fourier dimension). The *Fourier dimension* of $f\colon \mathbb{F}_2^n \to \{+1,-1\}$ denoted as $\mathit{dim}(f)$ is the smallest integer k such that there exists $A \leq \mathbb{F}_2^n$ of dimension k for which $\mathit{Spec}(f) \subseteq A$.

---

[11] In all Fourier-analytic arguments Boolean functions are treated as functions of the form $f : \mathbb{F}_2^n \to \{+1,-1\}$ where 0 is mapped to 1 and 1 is mapped to -1. Otherwise we use these two notations interchangeably.

We say that $A \leq \mathbb{F}_2^n$ is a *standard subspace* if it has a basis $v_1, \ldots, v_d$ where each v_i has Hamming weight equal to 1. An *orthogonal subspace* $A^\perp$ is defined as:

$$A^\perp = \{\gamma \in \mathbb{F}_2^n : \forall x \in A \quad \gamma \cdot x = 0\}.$$

An *affine subspace* (or coset) of $\mathbb{F}_2^n$ of the form $A = H + a$ for some $H \leq \mathbb{F}_2^n$ and $a \in \mathbb{F}_2^n$ is defined as:

$$A = \{\gamma \in \mathbb{F}_2^n : \forall x \in H^\perp \quad \gamma \cdot x = a \cdot x\}.$$

We now introduce notation for restrictions of functions to affine subspaces.

▶ **Definition 8.** Let $f : \mathbb{F}_2^n \to \mathbb{R}$ and $z \in \mathbb{F}_2^n$. We define $f^{+z} : \mathbb{F}_2^n \to \mathbb{R}$ as $f^{+z}(x) = f(x+z)$.

▶ Fact 9. The Fourier coefficients of f^{+z} are $\widehat{f^{+z}}(\gamma) = (-1)^{\gamma \cdot z}\hat{f}(\gamma)$ and hence:

$$f^{+z} = \sum_{S \in \mathbb{F}_2^n} \hat{f}(S)\chi_S(z)\chi_S.$$

▶ **Definition 10** (Coset restriction)**.** For $f : \mathbb{F}_2^n \to \mathbb{R}, z \in \mathbb{F}_2^n$ and $H \leq \mathbb{F}_2^n$ we write $f_H^{+z} : H \to \mathbb{R}$ for the restriction of f to $H + z$.

▶ **Definition 11** (Convolution)**.** For two functions $f, g : \mathbb{F}_2^n \to \mathbb{R}$ their convolution $(f * g) : \mathbb{F}_2^n \to \mathbb{R}$ is defined as $(f * g)(x) = \mathbb{E}_{y \sim U(\mathbb{F}_2^n)}[f(y)g(x+y)]$.

For $S \in \mathbb{F}_2^n$ the corresponding Fourier coefficient of convolution is given as $\widehat{f * g}(S) = \hat{f}(S)\hat{g}(S)$.

3 $\mathbb{F}_2$-sketching over the uniform distribution

We use the following definition of Fourier concentration that plays an important role in learning theory [27]. As mentioned above in all Fourier-analytic arguments we replace the range of the functions with $\{+1, -1\}$.

▶ **Definition 12** (Fourier concentration)**.** The spectrum of a function $f : \mathbb{F}_2^n \to \{+1, -1\}$ is ϵ-concentrated on a collection of Fourier coefficients $Z \subseteq \mathbb{F}_2^n$ if $\sum_{\alpha \in Z} \hat{f}^2(\alpha) \geq \epsilon$.

We now introduce the notion of *approximate Fourier dimension* of a Boolean function.

▶ **Definition 13** (Approximate Fourier dimension)**.** Let $\mathcal{A}_k$ be the set of all linear subspaces of $\mathbb{F}_2^n$ of dimension k. For $f : \mathbb{F}_2^n \to \{+1, -1\}$ and $\epsilon \in (0, 1]$ the ϵ-approximate Fourier dimension $\dim_\epsilon(f)$ is defined as:

$$\dim_\epsilon(f) = \min\left\{k : \exists A \in \mathcal{A}_k : \sum_{\alpha \in A} \hat{f}^2(\alpha) \geq \epsilon\right\}.$$

The following theorem shows that for uniformly distributed inputs, both the one-way communication complexity of f^+ and the linear sketch complexity of f are characterized by the approximate Fourier dimension of f. An immediate corollary is that, up to some slack in the dependence on the probability of error, the one-way communication complexity under the uniform distribution matches the linear sketch complexity. We note that the lower bounds given by this theorem are stronger than the basic extractor lower bound given in Appendix C.1. See Remark C.1 for further discussion.

▶ **Theorem 14.** *Let $f\colon \mathbb{F}_2^n \to \{+1,-1\}$ be a Boolean function. Let $\xi \in [0,1]$ and $\gamma < \frac{1-\sqrt{\xi}}{2}$. Let $d = \dim_\xi(f)$. Then,*

$$1.\ \mathcal{D}^{\to,U}_{(1-\xi)/2}(f^+) \le \mathcal{D}^{lin,U}_{(1-\xi)/2}(f) \le d, \qquad 2.\ \mathcal{D}^{lin,U}_{\gamma}(f) \ge d, \qquad 3.\ \mathcal{D}^{\to,U}_{(1-\xi)/6} \ge \frac{d}{6}.$$

Proof. Part 1[12]**.** Since $d = \dim_\xi(f)$, there exists a subspace $A \le \mathbb{F}_2^n$ of dimension at most d which satisfies $\sum_{\alpha\in A} \hat{f}^2(\alpha) \ge \xi$. Let $g\colon \mathbb{F}_2^n \to \mathbb{R}$ be a function defined by its Fourier transform as follows:

$$\hat{g}(\alpha) = \begin{cases} \hat{f}(\alpha), \text{ if } \alpha \in A \\ 0, \text{ otherwise.} \end{cases}$$

Consider drawing a random variable θ from the distribution with p.d.f $1 - |\theta|$ over $[-1,1]$.

▶ Proposition 15. For all t such that $-1 \le t \le 1$ and $z \in \{+1,-1\}$ random variable θ satisfies:

$$\Pr_\theta[sgn(t-\theta) \ne z] \le \frac{1}{2}(z-t)^2.$$

Proof. W.l.o.g we can assume $z = 1$ as the case $z = -1$ is symmetric. Then we have:

$$\Pr_\theta[sgn(t-\theta) \ne 1] = \int_t^1 (1-|\gamma|)d\gamma \le \int_t^1 (1-\gamma)d\gamma = \frac{1}{2}(1-t)^2. \qquad \blacksquare$$

Define a family of functions $g_\theta : \mathbb{F}_2^n \to \{+1,-1\}$ as $g_\theta(x) = sgn(g(x)-\theta)$. Then we have:

$$\begin{aligned}
\mathop{\mathbb{E}}_\theta\left[\Pr_{x\sim\mathbb{F}_2^n}[g_\theta(x) \ne f(x)]\right] &= \mathop{\mathbb{E}}_{x\sim\mathbb{F}_2^n}\left[\Pr_\theta[g_\theta(x) \ne f(x)]\right] \\
&= \mathop{\mathbb{E}}_{x\sim\mathbb{F}_2^n}\left[\Pr_\theta[sgn(g(x)-\theta) \ne f(x)]\right] \\
&\le \mathop{\mathbb{E}}_{x\sim\mathbb{F}_2^n}\left[\frac{1}{2}(f(x)-g(x))^2\right] \text{ (by Proposition 15)} \\
&= \frac{1}{2}\|f-g\|_2^2.
\end{aligned}$$

Using the definition of g and Parseval we have:

$$\frac{1}{2}\|f-g\|_2^2 = \frac{1}{2}\|\widehat{f-g}\|_2^2 = \frac{1}{2}\|\hat{f}-\hat{g}\|_2^2 = \frac{1}{2}\sum_{\alpha\notin A} \hat{f}^2(\alpha) \le \frac{1-\xi}{2}.$$

Thus, there exists a choice of θ such that g_θ achieves error at most $\frac{1-\xi}{2}$. Clearly g_θ can be computed based on the d parities forming a basis for A and hence $\mathcal{D}^{lin,U}_{(1-\xi)/2}(f) \le d$.

Part 2.

Fix any deterministic sketch that uses $d-1$ parities $\chi_{\alpha_1}, \ldots, \chi_{\alpha_{d-1}}$ and let $S = (\alpha_1, \ldots, \alpha_{d-1})$. For fixed values of these sketches $b = (b_1, \ldots, b_{d-1})$ where $b_i = \chi_{\alpha_i}(x)$ we denote the resulting

[12] This argument is a refinement of the standard "sign trick" from learning theory which approximates a Boolean function by taking a sign of its real-valued approximation under ℓ_2.

affine restriction of f as $f|_{(S,b)}$. Using the standard expression for the Fourier coefficients of an affine restriction the constant Fourier coefficient of the restricted function is given as:

$$\widehat{f|_{(S,b)}}(\emptyset) = \sum_{Z \subseteq [d-1]} (-1)^{\sum_{i \in Z} b_i} \hat{f}\left(\sum_{i \in Z} \alpha_i\right).$$

Thus, we have:

$$\widehat{f|_{(S,b)}}^2(\emptyset) = \sum_{Z \subseteq [d-1]} \hat{f}^2(\sum_{i \in Z} \alpha_i) + \sum_{Z_1 \neq Z_2 \subseteq [d-1]} (-1)^{\sum_{i \in Z_1 \Delta Z_2} b_i} \hat{f}(\sum_{i \in Z_1} \alpha_i)\hat{f}(\sum_{i \in Z_2} \alpha_i).$$

Taking expectation over a uniformly random $b \sim U(\mathbb{F}_2^d)$ we have:

$$\begin{aligned}
&\mathbb{E}_{b \sim U(\mathbb{F}_2^d)}\left[\widehat{f|_{(S,b)}}^2(\emptyset)\right] \\
=&\mathbb{E}_{b \sim U(\mathbb{F}_2^d)}\left[\sum_{Z \subseteq [d-1]} \hat{f}^2\left(\sum_{i \in Z} \alpha_i\right) + \right. \\
&\left. \sum_{Z_1 \neq Z_2 \subseteq [d-1]} (-1)^{\sum_{i \in Z_1 \Delta Z_2} b_i} \hat{f}\left(\sum_{i \in Z_1} \alpha_i\right) \hat{f}\left(\sum_{i \in Z_2} \alpha_i\right)\right] \\
=& \sum_{Z \subseteq [d-1]} \hat{f}^2\left(\sum_{i \in Z} \alpha_i\right).
\end{aligned}$$

The latter sum is the sum of squared Fourier coefficients over a linear subspace of dimension $d-1 < \dim_\xi(f)$, and hence is strictly less than ξ. Using Jensen's inequality:

$$\mathbb{E}_{b \sim U(\mathbb{F}_2^d)}\left[|\widehat{f|_{(S,b)}}(\emptyset)|\right] \leq \sqrt{\mathbb{E}_{b \sim U(\mathbb{F}_2^d)}\left[\widehat{f|_{(S,b)}}^2(\emptyset)\right]} < \sqrt{\xi}.$$

For a fixed restriction (S,b) if $|\hat{f}|_{(S,b)}(\emptyset)| < \alpha$ then $|\Pr[f|_{(S,b)} = 1] - \Pr[f|_{(S,b)} = -1]| < \alpha$ and hence no algorithm can predict the value of the restricted function on this coset with probability at least $\frac{1+\alpha}{2}$. Thus no algorithm can predict $f|_{(\alpha_1,b_1),\ldots,(\alpha_{d-1},b_{d-1})}$ for a uniformly random choice of $(b_1,\ldots,b_{d-1})$, and hence also on a uniformly at random chosen x, with probability at least $\frac{1+\sqrt{\xi}}{2}$.

Part 3.

We will need the following fact about entropy of a binary random variable. The proof is given in the appendix (Section A.1).

▶ Fact 16. For any random variable X supported on $\{1,-1\}$, $H(X) \leq 1 - \frac{1}{2}(\mathbb{E}X)^2$.

We will need the following proposition that states that random variables taking value in $\{1,-1\}$ that are highly biased have low variance. The proof of Proposition 17 can be found in the appendix (Section E.1).

▶ Proposition 17. Let X be a random variable taking values in $\{1,-1\}$. Define $p := \min_{b \in \{1,-1\}} \Pr[X = b]$. Then $\mathsf{Var}[X] \in [2p, 4p]$.

In the next two lemmas, we look into the structure of a one-way communication protocol for f^+, and analyze its performance when the inputs are uniformly distributed. We give a lower bound on the number of bits of information that any correct randomized one-way protocol reveals about Alice's input, in terms of the linear sketching complexity of f for uniform distribution[13].

The next lemma bounds the probability of error of a one-way protocol from below in terms of the Fourier coefficients of f, and the conditional distributions of different parities of Alice's input conditioned on Alice's random message.

▶ **Lemma 18.** *Let $\epsilon \in [0, \frac{1}{2})$. Let Π be a deterministic one-way protocol for f^+ such that $\Pr_{x,y \sim U(\mathbb{F}_2^n)}[\Pi(x,y) \neq f^+(x,y)] \leq \epsilon$. Let M denote the distribution of the random message sent by Alice to Bob in Π. For any fixed message m sent by Alice, let D_m denote the distribution of Alice's input x conditioned on the event that $M = m$. Then,*

$$4\epsilon \geq \sum_{\alpha \in \mathbb{F}_2^n} \widehat{f}^2(\alpha) \cdot \left(1 - \mathop{\mathbb{E}}_{m \sim M} \left(\mathop{\mathbb{E}}_{x \sim \mathsf{D}_m}[\chi_\alpha(x)]\right)^2\right).$$

Proof. For any fixed input y of Bob, define $\epsilon_m^{(y)} := \Pr_{x \sim \mathsf{D}_m}[\Pi(x,y) \neq f^+(x,y)]$. Thus,

$$\epsilon \geq \mathop{\mathbb{E}}_{m \sim M} \mathop{\mathbb{E}}_{y \sim U(\mathbb{F}_2^n)}[\epsilon_m^{(y)}]. \tag{1}$$

Note that the output of the protocol is determined by Alice's message and y. Hence for a fixed message and Bob's input, if the restricted function is largely unbiased, then any protocol is forced to commit an error with high probability. Formally,

$$\epsilon_m^{(y)} \geq \min_{b \in \{1,-1\}} \Pr_{x \sim \mathsf{D}_m}[f^+(x,y) = b] \geq \frac{\mathrm{Var}_{x \sim \mathsf{D}_m}[f^+(x,y)]}{4}. \tag{2}$$

Since $f^+(\cdot,\cdot)$ takes values in $\{+1,-1\}$, the second inequality follows from Proposition 17. Now,

$$\begin{aligned}
\mathrm{Var}_{x \sim \mathsf{D}_m}[f^+(x,y)] &= 1 - \left(\mathop{\mathbb{E}}_{x \sim \mathsf{D}_m}[f^+(x,y)]\right)^2 \qquad (\text{since } f^+(x,y) \in \{1,-1\}) \\
&= 1 - \left(\sum_{\alpha \in \mathbb{F}_2^n} \widehat{f}(\alpha)\chi_\alpha(y) \mathop{\mathbb{E}}_{x \sim \mathsf{D}_m}[\chi_\alpha(x)]\right)^2 \qquad (\text{by Fact 9 and linearity of expectation}) \\
&= 1 - \left(\sum_{\alpha \in \mathbb{F}_2^n} \widehat{f}^2(\alpha) \left(\mathop{\mathbb{E}}_{x \sim \mathsf{D}_m}[\chi_\alpha(x)]\right)^2\right. \\
&\qquad \left. + \sum_{(\alpha_1,\alpha_2) \in \mathbb{F}_2^n \times \mathbb{F}_2^n : \alpha_1 \neq \alpha_2} \widehat{f}(\alpha_1)\widehat{f}(\alpha_2)\chi_{\alpha_1+\alpha_2}(y) \mathop{\mathbb{E}}_{x \sim \mathsf{D}_m}[\chi_{\alpha_1}(x)] \mathop{\mathbb{E}}_{x \sim \mathsf{D}_m}[\chi_{\alpha_2}(x)]\right).
\end{aligned}$$

Taking expectation over y we have:

$$\mathop{\mathbb{E}}_{y \sim U(\mathbb{F}_2^n)}\left[\mathrm{Var}_{x \sim \mathsf{D}_m}[f^+(x,y)]\right] = 1 - \sum_{\alpha \in \mathbb{F}_2^n} \widehat{f}^2(\alpha) \left(\mathop{\mathbb{E}}_{x \sim \mathsf{D}_m}[\chi_\alpha(x)]\right)^2. \tag{3}$$

[13] We thus prove an *information complexity* lower bound. See, for example, [21] for an introduction to information complexity.

Taking expectation over messages it follows from (1), (2) and (3) that,

$$4\epsilon \geq 1 - \sum_{\alpha \in \mathbb{F}_2^n} \widehat{f}^2(\alpha) \cdot \mathop{\mathbb{E}}_{m \sim M} \left(\mathop{\mathbb{E}}_{x \sim \mathsf{D}_m} [\chi_\alpha(x)] \right)^2$$

$$= \sum_{\alpha \in \mathbb{F}_2^n} \widehat{f}^2(\alpha) \cdot \left(1 - \mathop{\mathbb{E}}_{m \sim M} \left(\mathop{\mathbb{E}}_{x \sim \mathsf{D}_m} [\chi_\alpha(x)] \right)^2 \right). \tag{4}$$

The second equality above follows from the Parseval's identity (Fact 6). The lemma follows. ∎

Let $\epsilon := \frac{1-\xi}{6}$. Let Π be a deterministic protocol such that $\Pr_{x,y \sim U(\mathbb{F}_2^n)}[\Pi(x,y) \neq f^+(x,y)] \leq \epsilon$, with optimal cost $c_\Pi := \mathcal{D}_\epsilon^{\rightarrow,U}(f^+) = \mathcal{D}_{\frac{1-\xi}{6}}^{\rightarrow,U}(f^+)$. Let M denote the distribution of the random message sent by Alice to Bob in Π. For any fixed message m sent by Alice, let D_m denote the distribution of Alice's input x conditioned on the event that $M = m$. To prove Part 3 of Theorem 14 we use the protocol Π to come up with a subspace of $\mathbb{F}_2^n$. Next, in Lemma 19 (a) we prove, using Lemma 18, that f is ξ-concentrated on that subspace. In Lemma 19 (b) we upper bound the dimension of that subspace in terms of c_Π.

▶ **Lemma 19.** *Let $\mathcal{A} := \{\alpha \in \mathbb{F}_2^n : \mathbb{E}_{m \sim M} \left(\mathbb{E}_{x \sim D_m} \chi_\alpha(x)\right)^2 \geq \frac{1}{3}\} \subseteq \mathbb{F}_2^n$. Let $\ell = \dim(\mathsf{span}(\mathcal{A}))$. Then,*

(a) $\ell \geq d$.

(b) $\ell \leq 6c_\Pi$.

Proof. (a) We prove part (a) by showing that f is ξ-concentrated on $\mathsf{span}(\mathcal{A})$. By Lemma 18 we have that

$$4\epsilon \geq \sum_{\alpha \in \mathsf{span}(\mathcal{A})} \widehat{f}^2(\alpha) \cdot \left(1 - \mathop{\mathbb{E}}_{m \sim M} \left(\mathop{\mathbb{E}}_{x \sim \mathsf{D}_m} \chi_\alpha(x) \right)^2 \right) +$$

$$\sum_{\alpha \notin \mathsf{span}(\mathcal{A})} \widehat{f}^2(\alpha) \cdot \left(1 - \mathop{\mathbb{E}}_{m \sim M} \left(\mathop{\mathbb{E}}_{x \sim \mathsf{D}_m} \chi_\alpha(x) \right)^2 \right)$$

$$> \frac{2}{3} \cdot \sum_{\alpha \notin \mathsf{span}(\mathcal{A})} \widehat{f}^2(\alpha).$$

Thus $\sum_{\alpha \notin \mathsf{span}(\mathcal{A})} \widehat{f}^2(\alpha) < 6\epsilon$. Hence, $\sum_{\alpha \in \mathsf{span}(\mathcal{A})} \widehat{f}^2(\alpha) \geq 1 - 6\epsilon = \xi$. Hence we have $\ell = \dim(\mathsf{span}(\mathcal{A})) \geq \dim_\xi(f) = d$.

(b) Notice that $\chi_\alpha(x)$ is a unbiased random variable taking values in $\{1, -1\}$. For each α in the set $\mathcal{A}$ in Proposition 19, the value of $\mathbb{E}_{m \sim M} \left(\mathbb{E}_{x \sim \mathsf{D}_m} \chi_\alpha(x)\right)^2$ is bounded away from 0. This suggests that for a typical message m drawn from M, the distribution of $\chi_\alpha(x)$ conditioned on the event $M = m$ is significantly biased. Fact 16 enables us to conclude that Alice's message reveals $\Omega(1)$ bit of information about $\chi_\alpha(x)$. However, since the total information content of Alice's message is at most c_Π, there can be at most $O(c_\Pi)$ independent vectors in $\mathcal{A}$. Now we formalize this intuition.

Let $\mathcal{T} = \{\alpha_1, \ldots, \alpha_\ell\}$ be a basis of $\mathsf{span}(\mathcal{A})$. Then,

$$
\begin{aligned}
c_\Pi &\geq H(M) && \text{(by the third inequality of Fact 5 (1))}\\
&\geq I(M; \chi_{\alpha_1}(x), \ldots, \chi_{\alpha_\ell}(x)) && \text{(by observation 7)}\\
&= H(\chi_{\alpha_1}(x), \ldots, \chi_{\alpha_\ell}(x)) - H(\chi_{\alpha_1}(x), \ldots, \chi_{\alpha_\ell}(x) \mid M)\\
&= \ell - H(\chi_{\alpha_1}(x), \ldots, \chi_{\alpha_\ell}(x) \mid M)\\
& \qquad \text{(by Fact 5 (3) as } \chi_{\alpha_i}(x)\text{'s are independent as random variables)}\\
&\geq \ell - \sum_{i=1}^{\ell} H(\chi_{\alpha_i}(x) \mid M) && \text{(by Fact 5 (2))}\\
&\geq \ell - \ell\left(1 - \frac{1}{2} \cdot \frac{1}{3}\right) && \text{(by Fact 16)}\\
&= \frac{\ell}{6}.
\end{aligned}
$$

∎

Recall that $c_\Pi = \mathcal{D}^{\rightarrow, U}_{\frac{1-\xi}{6}}(f^+)$. Part 3 of Theorem 14 follows easily from Lemma 19:

$$
\begin{aligned}
\mathcal{D}^{\rightarrow, U}_{\frac{1-\xi}{6}}(f^+) &= c_\Pi\\
&\geq \frac{\ell}{6} && \text{(by Lemma 19 (b))}\\
&\geq \frac{d}{6}. && \text{(by Lemma 19 (a))}
\end{aligned}
$$

∎

The proof of Theorem 4 now follows directly from Part 1 and Part 3 of Theorem 14 by setting $\xi = 1/3$.

4 Applications

In this section using Theorem 14 we confirm Conjecture 3 for several funcion classes: low-degree $\mathbb{F}_2$ polynomials, functions with sparse Fourier spectrum and symmetric functions (which are not too imbalanced). We also give an example of a composition theorem using recursive majority function as an example.

4.1 Low-degree $\mathbb{F}_2$ polynomials

In this section we show that for Boolean functions with low $\mathbb{F}_2$-degree randomness does not help in the design of linear sketches or one-way communication protocols. We briefly review some basic definitions, facts and results below.

▶ **Fact 20.** For every Boolean function $f : \mathbb{F}_2^n \to \mathbb{F}_2$ there is a unique n-variate polynomial $p \in \mathbb{F}_2[x_1, \ldots, x_n]$ such that for every $(x_1, \ldots, x_n) \in \mathbb{F}_2^n$, $f(x_1, \ldots, x_n) = p(x_1, \ldots, x_n)$.

The uniqueness of this representation in particular implies that the only $\mathbb{F}_2$ polynomial representing the constant 0 function is the polynomial 0. Taking the contrapositive, we have that for every non-constant $\mathbb{F}_2$ polynomial there is an assignment to its input variables on which the polynomial evaluates to 1.

The degree of p is referred to as the $\mathbb{F}_2$-degree of f. We will need the following standard result which states that a function with low $\mathbb{F}_2$-degree cannot vanish on too many points in its domain. For the sake of completion, we add a proof of it in the appendix (Section E.2).

▶ **Lemma 21.** *Let f be a Boolean function different than the constant 0 function with $\mathbb{F}_2$ degree d. Then,*

$$\Pr_{x \sim U(\mathbb{F}_2^n)}[f(x) = 1] \geq \frac{1}{2^d}.$$

In this section we prove the following theorem.

▶ **Theorem 22.** *Let $f : \mathbb{F}_2^n \to \mathbb{F}_2$ be a Boolean function, and let the $\mathbb{F}_2$-degree of f be d. Then,*

$$D^{lin}(f) = \dim(f) = O\left(R^{\to}_{1/3}(f^+) \cdot d\right).$$

Proof. Let $\ell = \mathcal{D}^{lin,U}_{\frac{1}{4 \cdot 2^d}}(f)$. This implies that there is a set $\mathcal{P} = \{P_1, \ldots, P_\ell\}$ of at most ℓ parities and a Boolean function g such that $\Pr_{x \sim U(\mathbb{F}_2^n)}[f(x) \neq g(P_1(x), \ldots, P_\ell(x))] \leq \frac{1}{4 \cdot 2^d}$. We now prove that $D^{lin}(f)$ (or equivalently Fourier dimension) of f is at most ℓ. That will prove the theorem as:

$$\mathcal{D}^{lin,U}_{\frac{1}{4 \cdot 2^d}}(f) = O\left(\mathcal{D}^{\to,U}_{\frac{1}{12 \cdot 2^d}}(f^+)\right),$$
$$\mathcal{D}^{\to,U}_{\frac{1}{12 \cdot 2^d}}(f^+) = O\left(R^{\to}_{\frac{1}{12 \cdot 2^d}}(f^+)\right),$$
$$R^{\to}_{\frac{1}{12 \cdot 2^d}}(f^+) = O\left(R^{\to}_{1/3}(f^+) \cdot d\right).$$

where the first relation follows by invoking parts 1 and 3 of Theorem 14 with $\xi = 1 - \frac{1}{2^{d+1}}$, the second relation holds by fixing the randomness of a randomized one-way protocol appropriately, and the third relation is true because the error of a randomized one-way protocol can be reduced from $1/3$ to $\frac{1}{12 \cdot 2^d}$ by taking the majority of $O(d)$ independent parallel repetitions.

It is left to prove that $D^{lin}(f) \leq \ell$. We prove it by showing that evaluations of all the parities in the set $\mathcal{P}$ determine the value of f. For each $b = (b_1, \ldots, b_\ell) \in \mathbb{F}_2^\ell$, let V_b denote the affine subspace $\{x \in \mathbb{F}_2^n : P_1(x) = b_1, \ldots, P_\ell(x) = b_\ell\}$ and define:

$$p_b := \Pr_{x \sim U(V_b)}[f(x) \neq g(P_1(x), \ldots, P_\ell(x))] = \Pr_{x \sim U(V_b)}[f(x) \neq g(b_1, \ldots, b_\ell)].$$

Note that:

$$p_b \geq \min\{\Pr_{x \sim U(V_b)}[f(x) = 0], \Pr_{x \sim U(V_b)}[f(x) = 1]\} \geq \frac{1}{2} \Pr_{x, x' \sim U(V_b)}[f(x) \neq f(x')]. \tag{5}$$

Given this observation, define $F : \mathbb{F}_2^n \times \mathbb{F}_2^n \to \mathbb{F}_2$ as follows. For $x, x' \in \mathbb{F}_2^n$ let:

$$F(x, x') := \mathbf{1}_{f(x) \neq f(x')} = f(x) + f(x') \mod 2.$$

Note that $\mathbb{F}_2$-degree of F is at most d. Now,

$$\begin{aligned}
& \Pr_{x\sim U(\mathbb{F}_2^n)}[f(x) \neq g(P_1(x),\dots,P_\ell(x))] \leq \frac{1}{4\cdot 2^d} \\
\Rightarrow\quad & \mathbb{E}_{b\sim U(\mathbb{F}_2^\ell)}\left[\Pr_{x\sim U(V_b)}[f(x)\neq g(b_1,\dots,b_\ell)]\right] \leq \frac{1}{4\cdot 2^d} \\
\Rightarrow\quad & \mathbb{E}_{b\sim U(\mathbb{F}_2^\ell)}[p_b] \leq \frac{1}{4\cdot 2^d} \\
\Rightarrow\quad & \mathbb{E}_{b\sim U(\mathbb{F}_2^\ell)}\left[\Pr_{x,x'\sim U(V_b)}[f(x)\neq f(x')]\right] \leq \frac{1}{2\cdot 2^d} \quad \text{(From equation (5))} \\
\Rightarrow\quad & \mathbb{E}_{b\sim U(\mathbb{F}_2^\ell)}\left[\Pr_{x,x'\sim U(V_b)}[F(x,x')=1]\right] \leq \frac{1}{2\cdot 2^d} \qquad (6)
\end{aligned}$$

Let V denote the subspace $\{(x,x') \in \mathbb{F}_2^n \times \mathbb{F}_2^n : P_1(x) = P_1(x'),\dots,P_\ell(x) = P_\ell(x')\}$ of $\mathbb{F}_2^n \times \mathbb{F}_2^n$. From 6 we have that

$$\Pr_{(x,x')\sim U(V)}[F(x,x')=1] \leq \frac{1}{2\cdot 2^d} < \frac{1}{2^d}. \qquad (7)$$

Since $\mathbb{F}_2$-degree of F is at most d, restriction of F to V also has $\mathbb{F}_2$ degree at most d. Equation 7 and Fact 21 imply that F is the constant 0 function on V. Thus for each x, x' such that $P_1(x) = P_1(x'),\dots,P_\ell(x) = P_\ell(x')$, $f(x) = f(x')$. Thus $f(x)$ is a function of $P_1(x),\dots,P_\ell(x)$. Hence, Fourier dimension of f is at most ℓ. ∎

For low-degree polynomials with bounded spectral norm we obtain the following corollary.

▶ **Corollary 23.** *Let $f : \mathbb{F}_2^n \to \mathbb{F}_2$ be a Boolean function of $\mathbb{F}_2$-degree d. Then*

$$D^{lin}(f) = dim(f) = O\left(d\cdot \|\hat{f}\|_1^2\right).$$

Proof. The proof follows from the result of Grolmusz [17, 39] that shows that $R^{\rightarrow}_{1/3}(f^+) = O(\|\hat{f}\|_1^2)$ and Theorem 22. ∎

This result should be compared with Corollary 6 in Tsang et al. [49] who show that $D^{lin}(f) = O(2^{d^3/2}\log^{d^2}\|\hat{f}\|_1)$. Corollary 23 gives a stronger bound for $d = \omega\left(\log^{1/3}\|\hat{f}\|_1\right)$.

4.2 Address function and Fourier sparsity

Consider the *addressing function* $Add_n\colon \{0,1\}^{\log n + n} \to \{0,1\}$ defined as follows[14]:

$$Add_n(x, y_1,\dots,y_n) = y_x, \text{ where } x \in \{0,1\}^{\log n}, y_i \in \{0,1\},$$

i.e. the value of Add_n on an input (x,y) is given by the x-th bit of the vector y where x is treated as a binary representation of an integer number in between 1 and n. Here x is commonly referred to as the *address block* and y as the *addressee block*. Addressing function has only n^2 non-zero Fourier coefficients. In fact, as shown by Sanyal [44] the Fourier dimension, and hence by Fact 8 also the deterministic sketch complexity, of any Boolean function with Fourier sparsity s is $O(\sqrt{s}\log s)$.

Below using the addressing function we show that this relationship is tight (up to a logarithmic factor) even if randomization is allowed, i.e. even for a function with Fourier sparsity s an $\mathbb{F}_2$ sketch of size $\Omega(\sqrt{s})$ might be required.

[14] In this section it will be more convenient to represent both domain and range of the function using $\{0,1\}$ rather than $\mathbb{F}_2$.

▶ **Theorem 24.** *For the addressing function Add_n and values $1 \le d \le n$ and $\xi > d/n$ it holds that:*

$$\mathcal{D}^{lin,U}_{\frac{1-\sqrt{\xi}}{2}}(Add_n^+) > d, \qquad \mathcal{D}^{\to,U}_{\frac{1-\xi}{6}}(Add_n) > \frac{d}{6}.$$

Proof. If we apply the standard Fourier notation switch where we replace 0 with 1 and 1 with -1 in the domain and the range of the function then the addressing function $Add_n(x, y)$ can be expressed as the following multilinear polynomial:

$$Add_n(x,y) = \sum_{i \in \{0,1\}^{\log n}} y_i \prod_{j\,:\,i_j=1} \left(\frac{1-x_j}{2}\right) \prod_{j\,:\,i_j=0} \left(\frac{1+x_j}{2}\right),$$

which makes it clear that the only non-zero Fourier coefficents correspond to the sets that contain a single variable from the addressee block and an arbitrary subset of variables from the address block. This expansion also shows that the absolute value of each Fourier coefficient is equal to $\frac{1}{n}$.

Fix any d-dimensional subspace $\mathcal{A}_d$ and consider the matrix $M \in \mathbb{F}_2^{d\times(\log n+n)}$ composed of the basis vectors as rows. We add to M extra $\log n$ rows which contain an identity matrix in the first $\log n$ coordinates and zeros everywhere else. This gives us a new matrix $M' \in \mathbb{F}_2^{(d+\log n)\times(\log n+n)}$. Applying Gaussian elimination to M' we can assume that it is of the following form:

$$M' = \begin{pmatrix} I_{\log n} & 0 & 0 \\ 0 & I_{d'} & M'' \\ 0 & 0 & 0 \end{pmatrix},$$

where $d' \le d$. Thus, the total number of non-zero Fourier coefficients spanned by the rows of M' equals nd'. Hence, the total sum of squared Fourier coeffients in $\mathcal{A}_d$ is at most $\frac{d'}{n} \le \frac{d}{n}$, i.e. $\dim_\xi(Add_n) > d$. By Part 2 and Part 3 of Theorem 14 the statement of the theorem follows. ∎

4.3 Symmetric functions

A function $f : \mathbb{F}_2^n \to \mathbb{F}_2$ is symmetric if it can be expressed as $g(\|x\|_0)$ for some function $g : [0, n] \to \mathbb{F}_2$. We give the following lower bound for symmetric functions:

▶ **Theorem 25** (Lower bound for symmetric functions)**.** *For any symmetric function $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$ that isn't $(1-\epsilon)$-concentrated on $\{\emptyset, \{1, \dots, n\}\}$:*

$$\mathcal{D}^{lin,U}_{\epsilon/8}(f) \ge \frac{n}{2e}, \qquad \mathcal{D}^{\to,U}_{\epsilon/12}(f^+) \ge \frac{n}{2e}.$$

Proof. First we prove an auxiliary lemma. Let W_k be the set of all vectors in $\mathbb{F}_2^n$ of Hamming weight k.

▶ **Lemma 26.** *For any $d \in [n/2]$, $k \in [n-1]$ and any d-dimensional subspace $\mathcal{A}_d \le \mathbb{F}_2^n$:*

$$\frac{|W_k \cap \mathcal{A}_d|}{|W_k|} \le \left(\frac{ed}{n}\right)^{min(k,n-k,d)} \le \frac{ed}{n}.$$

Proof. Fix any basis in $\mathcal{A}_d$ and consider the matrix $M \in \mathbb{F}_2^{d\times n}$ composed of the basis vectors as rows. W.l.o.g we can assume that this matrix is diagonalized and is in the standard form (I_d, M') where I_d is a $d \times d$ identity matrix and M' is a $d \times (n-d)$-matrix. Clearly, any linear combination of more than k rows of M has Hamming weight greater than k just from the contribution of the first d coordinates. Thus, we have $|W_k \cap \mathcal{A}_d| \le \sum_{i=0}^{k} \binom{d}{i}$.

For any $k \le d$ it is a standard fact about binomials that $\sum_{i=0}^{k} \binom{d}{i} \le \left(\frac{ed}{k}\right)^k$. On the other hand, we have $|W_k| = \binom{n}{k} \ge (n/k)^k$. Thus, we have $\frac{|W_k \cap \mathcal{A}_d|}{|W_k|} \le \left(\frac{ed}{n}\right)^k$ and hence for $1 \le k \le d$ the desired inequality holds.

If $d < k$ then consider two cases. Since $d \le n/2$ the case $n - d \le k \le n-1$ is symmetric to $1 \le k \le d$. If $d < k < n - d$ then we have $|W_k| > |W_d| \ge (n/d)^d$ and $|W_k \cap \mathcal{A}_d| \le 2^d$ so that the desired inequality follows. ∎

Any symmetric function has its spectrum distributed uniformly over Fourier coefficients of any fixed weight. Let $w_i = \sum_{S \in W_i} \hat{f}^2(S)$. By the assumption of the theorem we have $\sum_{i=1}^{n-1} w_i \ge \epsilon$. Thus, by Lemma 26 any linear subspace $\mathcal{A}_d$ of dimension at most $d \le n/2$ satisfies that:

$$\begin{aligned}\sum_{S \in \mathcal{A}_d} f^2(S) &\le \hat{f}^2(\emptyset) + \hat{f}^2(\{1, \dots, n\}) + \sum_{i=1}^{n-1} w_i \frac{|W_i \cap \mathcal{A}_d|}{|W_i|} \\ &\le \hat{f}^2(\emptyset) + \hat{f}^2(\{1, \dots, n\}) + \sum_{i=1}^{n-1} w_i \frac{ed}{n} \\ &\le (1 - \epsilon) + \epsilon \frac{ed}{n}.\end{aligned}$$

Thus, f isn't $1 - \epsilon(1 - \frac{ed}{n})$-concentrated on any d-dimensional linear subspace, i.e. $\dim_\xi(f) > d$ for $\xi = 1 - \epsilon(1 - \frac{ed}{n})$. By Part 2 of Theorem 14 this implies that f doesn't have randomized sketches of dimension at most d which err with probability less than:

$$\frac{1}{2} - \frac{\sqrt{1 - \epsilon(1 - \frac{ed}{n})}}{2} \ge \frac{\epsilon}{4}\left(1 - \frac{ed}{n}\right) \ge \frac{\epsilon}{8}$$

where the last inequality follows by the assumption that $d \le \frac{n}{2e}$. The communication complexity lower bound follows by Part 3 of Theorem 14 by setting $d = \frac{n}{2e}$.

4.4 Composition theorem for majority

In this section using Theorem 14 we give a composition theorem for $\mathbb{F}_2$-sketching of the composed Maj_3 function. Unlike in the deterministic case for which the composition theorem is easy to show (see Lemma 13) in the randomized case composition results require more work.

▶ **Definition 27** (Composition). For $f \colon \mathbb{F}_2^n \to \mathbb{F}_2$ and $g \colon \mathbb{F}_2^m \to \mathbb{F}_2$ their composition $f \circ g \colon \mathbb{F}_2^{mn} \to \mathbb{F}_2$ is defined as:

$$(f \circ g)(x) = f(g(x_1, \dots, x_m), g(x_{m+1}, \dots, x_{2m}), \dots, g(x_{m(n-1)+1}, \dots, x_{mn})).$$

Consider the recursive majority function $Maj_3^{\circ k} \equiv Maj_3 \circ Maj_3 \circ \cdots \circ Maj_3$ where the composition is taken k times.

▶ **Theorem 28.** *For any $d \le n$, $k = \log_3 n$ and $\xi > \frac{4d}{n}$ it holds that $\dim_\xi \left(Maj_3^{\circ k}\right) > d$.*

First, we show a slighthly stronger result for standard subspaces and then extend this result to arbitrary subspaces with a loss of a constant factor. Fix any set $S \subseteq [n]$ of variables. We associate this set with a collection of standard unit vectors corresponding to these variables. Hence in this notation $\emptyset$ corresponds to the all-zero vector.

▶ **Lemma 29.** *For any standard subspace whose basis consists of singletons from the set $S \subseteq [n]$ it holds that:*

$$\sum_{Z \in span(S)} \left(\widehat{Maj_3^{\circ k}}(Z)\right)^2 \leq \frac{|S|}{n}$$

Proof. The Fourier expansion of Maj_3 is given as

$$Maj_3(x_1, x_2, x_3) = \frac{1}{2}(x_1 + x_2 + x_3 - x_1x_2x_3)$$

. For $i \in \{1, 2, 3\}$ let $N_i = \{(i-1)n/3 + 1, \ldots, in/3\}$. Let $S_i = S \cap N_i$. Let α_i be defined as:

$$\alpha_i = \sum_{Z \in span(S_i)} \left(\widehat{Maj_3^{\circ k-1}}(Z)\right)^2.$$

Then we have:

$$\sum_{Z \in span(S)} \left(\widehat{Maj_3^{\circ k}}(Z)\right)^2 = \sum_{i=1}^{3} \sum_{Z \in span(S_i)} \left(\widehat{Maj_3^{\circ k}}(Z)\right)^2 + \sum_{Z \in span(S) - \cup_{i=1}^3 span(S_i)} \left(\widehat{Maj_3^{\circ k}}(Z)\right)^2.$$

For each S_i we have

$$\sum_{Z \in span(S_i)} \left(\widehat{Maj_3^{\circ k}}(Z)\right)^2 = \frac{1}{4} \sum_{Z \in span(S_i)} \left(\widehat{Maj_3^{\circ k-1}}(Z)\right)^2 = \frac{\alpha_i}{4}.$$

Moreover, for each $Z \in span(S) - \cup_{i=1}^3 span(S_i)$ we have:

$$\widehat{Maj_3^{\circ k}}(Z) = \begin{cases} -\frac{1}{2}\widehat{Maj_3^{\circ k-1}}(Z_1)\widehat{Maj_3^{\circ k-1}}(Z_2)\widehat{Maj_3^{\circ k-1}}(Z_3) & \text{if } Z \in \times_{i=1}^3 (span(S_i) \setminus \emptyset) \\ 0 & \text{otherwise.} \end{cases}$$

Thus, we have:

$$\begin{aligned}
&\sum_{Z \in (span(S_1)\setminus\emptyset)\times(span(S_2)\setminus\emptyset)\times(span(S_3)\setminus\emptyset)} \left(\widehat{Maj_3^{\circ k}}(Z)\right)^2 \\
&= \sum_{Z \in (span(S_1)\setminus\emptyset)\times(span(S_2)\setminus\emptyset)\times(span(S_3)\setminus\emptyset)} \frac{1}{4}\left(\widehat{Maj_3^{\circ k-1}}(Z_1)\right)^2 \cdot \left(\widehat{Maj_3^{\circ k-1}}(Z_2)\right)^2 \cdot \left(\widehat{Maj_3^{\circ k-1}}(Z_3)\right)^2 \\
&= \frac{1}{4}\left(\sum_{Z \in (span(S_1)\setminus\emptyset)} \left(\widehat{Maj_3^{\circ k-1}}(Z_1)\right)^2\right) \cdot \left(\sum_{Z \in (span(S_2)\setminus\emptyset)} \left(\widehat{Maj_3^{\circ k-1}}(Z_2)\right)^2\right) \cdot \left(\sum_{Z \in (span(S_3)\setminus\emptyset)} \left(\widehat{Maj_3^{\circ k-1}}(Z_3)\right)^2\right) \\
&= \frac{1}{4}\alpha_1\alpha_2\alpha_3.
\end{aligned}$$

where the last equality holds since $\widehat{Maj_3^{\circ k-1}}(\emptyset) = 0$. Putting this together we have:

$$\sum_{Z \in span(S)} \left(\widehat{Maj_3^{\circ k}}(Z)\right)^2 = \frac{1}{4}(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_1\alpha_2\alpha_3)$$

$$\le \frac{1}{4}\left(\alpha_1 + \alpha_2 + \alpha_3 + \frac{1}{3}(\alpha_1 + \alpha_2 + \alpha_3)\right) = \frac{1}{3}(\alpha_1 + \alpha_2 + \alpha_3).$$

Applying this argument recursively to each α_i for $k-1$ times we have:

$$\sum_{Z \in span(S)} \left(\widehat{Maj_3^{\circ k}}(Z)\right)^2 \le \frac{1}{3^k}\sum_{i=1}^{3^k} \gamma_i,$$

where $\gamma_i = 1$ if $i \in S$ and 0 otherwise. Thus, $\sum_{Z \in span(S)} \left(\widehat{Maj_3^{\circ k}}(Z)\right)^2 \le \frac{|S|}{n}$. ∎

To extend the argument to arbitrary linear subspaces we show that any such subspace has less Fourier weight than a collection of three carefully chosen standard subspaces. First we show how to construct such subspaces in Lemma 30.

For a linear subspace $L \le \mathbb{F}_2^n$ we denote the set of all vectors in L of odd Hamming weight as $\mathcal{O}(L)$ and refer to it as the *odd set* of L. For two vectors $v_1, v_2 \in \mathbb{F}_2^n$ we say that v_1 *dominates* v_2 if the set of non-zero coordinates of v_1 is a (not necessarily proper) subset of the set of non-zero coordinates of v_2. For two sets of vectors $S_1, S_2 \subseteq \mathbb{F}_2^n$ we say that S_1 *dominates* S_2 (denoted as $S_1 \prec S_2$) if there is a matching M between S_1 and S_2 of size $|S_2|$ such that for each $(v_1 \in S_1, v_2 \in S_2) \in M$ the vector v_1 dominates v_2.

▶ **Lemma 30** (Standard subspace domination lemma). *For any linear subspace $L \le \mathbb{F}_2^n$ of dimension d there exist three standard linear subspaces $S_1, S_2, S_3 \le \mathbb{F}_2^n$ such that:*

$$\mathcal{O}(L) \prec \mathcal{O}(S_1) \cup \mathcal{O}(S_2) \cup \mathcal{O}(S_3),$$

and $dim(S_1) = d-1$, $dim(S_2) = d$, $dim(S_3) = 2d$.

Proof. Let $A \in \mathbb{F}_2^{d \times n}$ be the matrix with rows corresponding to the basis in L. We will assume that A is normalized in a way described below. First, we apply Gaussian elimination to ensure that $A = (I, M)$ where I is a $d \times d$ identity matrix. If all rows of A have even Hamming weight then the lemma holds trivially since $\mathcal{O}(L) = \emptyset$. By reordering rows and columns of A we can always assume that for some $k \ge 1$ the first k rows of A have odd Hamming weight and the last $d-k$ have even Hamming weight. Finally, we add the first column to each of the last $d-k$ rows, which makes all rows have odd Hamming weight. This results in A of the following form:

$$A = \left(\begin{array}{c|c|c|c}
1 & 0\cdots0 & 0\cdots0 & a \\
\hline
\begin{matrix}0\\ \vdots \\ 0\end{matrix} & I_{k-1} & 0 & M_1 \\
\hline
\begin{matrix}1\\ \vdots \\ 1\end{matrix} & 0 & I_{d-k} & M_2
\end{array}\right)$$

We use the following notation for submatrices: $A[i_1, j_1; i_2, j_2]$ refers to the submatrix of A with rows between i_1 and j_1 and columns between i_2 and j_2 inclusive. We denote to the first row by v, the submatrix $A[2,k;1,n]$ as $\mathcal{A}$ and the submatrix $A[k+1,d;1,n]$ as $\mathcal{B}$. Each $x \in \mathcal{O}(L)$ can be represented as $\sum_{i \in S} A_i$ where the set S is of odd size and the sum is over $\mathbb{F}_2^n$. We consider the following three cases corresponding to different types of the set S.

Case 1. $S \subseteq rows(\mathcal{A}) \cup rows(\mathcal{B})$. This corresponds to all odd size linear combinations of the rows of A that don't include the first row. Clearly, the set of such vectors is dominated by $\mathcal{O}(S_1)$ where S_1 is the standard subspace corresponding to the span of the rows of the submatrix $A[2,d;2,d]$.

Case 2. S contains the first row, $|S \cap rows(\mathcal{A})|$ and $|S \cap rows(\mathcal{B})|$ are even. All such linear combinations have their first coordinate equal 1. Hence, they are dominated by a standard subspace corresponding to span of the rows the $d \times d$ identity matrix, which we refer to as S_2.

Case 3. S contains the first row, $|S \cap rows(\mathcal{A})|$ and $|S \cap rows(\mathcal{B})|$ are odd. All such linear combinations have their first coordinate equal 0. This implies that the Hamming weight of the first d coordinates of such linear combinations is even and hence the other coordinates cannot be all equal to 0. Consider the submatrix $M = A[1,d;d+1,n]$ corresponding to the last $n-d$ columns of A. Since the rank of this matrix is at most d by running Gaussian elimination on M we can construct a matrix M' containing as rows the basis for the row space of M of the following form:

$$M' = \begin{pmatrix} I_t & M_1 \\ 0 & 0 \end{pmatrix}$$

where $t = rank(M)$. This implies that any non-trivial linear combination of the rows of M contains 1 in one of the first t coordinates. We can reorder the columns of A in such a way that these t coordinates have indices from $d+1$ to $d+t$. Note that now the set of vectors spanned by the rows of the $(d+t) \times (d+t)$ identity matrix I_{d+t} dominates the set of linear combinations we are interested in. Indeed, each such linear combination has even Hamming weight in the first d coordinates and has at least one coordinate equal to 1 in the set $\{d+1, \dots, d+t\}$. This gives a vector of odd Hamming weight that dominates such linear combination. Since this mapping is injective we have a matching. We denote the standard linear subspace constructed this way by S_3 and clearly $dim(S_3) \leq 2d$. ∎

The following proposition shows that the spectrum of the $Maj_3^{\circ k}$ is monotone decreasing under inclusion if restricted to odd size sets only:

▶ Proposition 31. For any two sets $Z_1 \subseteq Z_2$ of odd size it holds that:

$$\left|\widehat{Maj_3^{\circ k}}(Z_1)\right| \geq \left|\widehat{Maj_3^{\circ k}}(Z_2)\right|.$$

Proof. The proof is by induction on k. Consider the Fourier expansion of $Maj_3(x_1, x_2, x_3) = \frac{1}{2}(x_1 + x_2 + x_3 - x_1x_2x_3)$. The case $k = 1$ holds since all Fourier coefficients have absolute value $1/2$. Since $Maj_3^{\circ k} = Maj_3 \circ (Maj_3^{\circ k-1})$ all Fourier coefficients of $Maj_3^{\circ k}$ result from substituting either a linear or a cubic term in the Fourier expansion by the multilinear expansions of $Maj_3^{\circ k-1}$. This leads to four cases.

Case 1. Z_1 and Z_2 both arise from linear terms. In this case if Z_1 and Z_2 aren't disjoint then they arise from the same linear term and thus satisfy the statement by the inductive hypothesis.

Case 2. If Z_1 arises from a cubic term and Z_2 from the linear term then it can't be the case that $Z_1 \subseteq Z_2$ since Z_2 contains some variables not present in Z_1.

Case 3. If Z_1 and Z_2 both arise from the cubic term then we have $(Z_1 \cap N_i) \subseteq (Z_2 \cap N_i)$ for each i. By the inductive hypothesis we then have $\left|\widehat{Maj_3^{\circ k-1}}(Z_1 \cap N_i)\right| \geq \left|\widehat{Maj_3^{\circ k-1}}(Z_2 \cap N_i)\right|$.

Since for $j = 1, 2$ we have $\widehat{Maj_3^{\circ k}}(Z_j) = -\frac{1}{2}\prod_i \widehat{Maj_3^{\circ k-1}}(Z_j \cap N_i)$ the desired inequality follows.

Case 4. If Z_1 arises from the linear term and Z_2 from the cubic term then w.l.o.g. assume that Z_1 arises from the x_1 term. Note that $Z_1 \subseteq (Z_2 \cap N_1)$ since $Z_1 \cap (N_2 \cup N_3) = \emptyset$. By the inductive hypothesis applied to Z_1 and $Z_2 \cap N_1$ the desired inequality holds.

We can now complete the proof of Theorem 28

Proof of Theorem 28. By combining Proposition 31 and Lemma 29 we have that any set $\mathcal{T}$ of vectors that is dominated by $\mathcal{O}(\mathcal{S})$ for some standard subspace $\mathcal{S}$ satisfies $\sum_{S\in\mathcal{T}} \widehat{Maj_3^{\circ k}}(S)^2 \leq \frac{dim(\mathcal{S})}{n}$. By the standard subspace domination lemma (Lemma 30) any subspace $L \leq \mathbb{F}_2^n$ of dimension d has $\mathcal{O}(L)$ dominated by a union of three standard subspaces of dimension $2d$, d and $d-1$ respectively. Thus, we have $\sum_{S\in\mathcal{O}(L)} \widehat{Maj_3^{\circ k}}(S)^2 \leq \frac{2d}{n} + \frac{d}{n} + \frac{d-1}{n} \leq \frac{4d}{n}$. ∎

We have the following corollary of Theorem 28 that proves Theorem 5.

▶ **Corollary 32.** *For any $\epsilon \in [0, \frac{1}{2}]$, $\xi > 4\epsilon^2$ and $k = \log_3 n$ it holds that:*

$$\mathcal{D}^{lin,U}_{\frac{1-\sqrt{\xi}}{2}}(Maj_3^{\circ k}) > \epsilon^2 n, \qquad \mathcal{D}^{\rightarrow,U}_{\frac{1-\xi}{6}}(Maj_3^{\circ k^+}) > \frac{\epsilon^2 n}{6}.$$

Proof. Fix $d = \epsilon^2 n$. For this choice of d Theorem 28 implies that for $\xi > 4\epsilon^2$ it holds tha t $\dim_\xi\left(Maj_3^{\circ k}\right) > d$. The first part follows from Part 2 of Theorem 14. The second part is by Part 3 of Theorem 14.

5 Streaming algorithms over $\mathbb{F}_2$

Let e_i be the standard unit vector in $\mathbb{F}_2^n$. In the turnstile streaming model the input $x \in \mathbb{F}_2^n$ is represented as a stream $\sigma = (\sigma_1, \sigma_2, \ldots)$ where $\sigma_i \in \{e_1, \ldots, e_n\}$. For a stream σ the resulting vector x corresponds to its frequency vector $\mathrm{freq}\ \sigma \equiv \sum_i \sigma_i$. Concatenation of two streams σ and τ is denoted as $\sigma \circ \tau$.

5.1 Random streams

In this section we show how to translate our results in Section 3 and 4 into lower bounds for streaming algorithms. We consider the following two natural models of random streams over $\mathbb{F}_2$:

Model 1. In the first model we start with $x \in \mathbb{F}_2^n$ that is drawn from the uniform distribution over $\mathbb{F}_2^n$ and then apply a uniformly random update $y \sim U(\mathbb{F}_2^n)$ obtaining $x + y$. In the streaming language this corresponds to a stream $\sigma = \sigma_1 \circ \sigma_2$ where $\mathrm{freq}\ \sigma_1 \sim U(\mathbb{F}_2^n)$ and $\mathrm{freq}\ \sigma_2 \sim U(\mathbb{F}_2^n)$. A specific example of such stream would be one where for both σ_1 and σ_2 we flip an unbiased coin to decide whether or not to include a vector e_i in the stream for each value of i. The expected length of the stream in this case is n.

Model 2. In the second model we consider a stream σ which consists of uniformly random updates. Let $\sigma_i = e_{r(i)}$ where $r(i) \sim U([n])$. This corresponds to each update being a flip in a coordinate of x chosen uniformly at random. This model is equivalent to the previous model but requires longer streams to mix. Using coupon collector's argument such streams of length $\Theta(n \log n)$ can be divided into two substreams σ_1 and σ_2 such that with high probability both $\mathrm{freq}\ \sigma_1$ and $\mathrm{freq}\ \sigma_2$ are uniformly distributed over $\mathbb{F}_2^n$ and $\sigma = \sigma_1 \circ \sigma_2$.

▶ **Theorem 33.** *Let $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$ be an arbitrary function. In the two random streaming models for generating σ described above any algorithm that computes $f(freq\ \sigma)$ with probability at least $8/9$ in the end of the stream has to use space that is at least $\mathcal{D}^{lin,U}_{1/3}(f)$.*

Proof. The proof follows directly from Theorem 4 as in both models we can partition the stream into σ_1 and σ_2 such that freq σ_1 and freq σ_2 are both distributed uniformly over $\mathbb{F}_2^n$. We treat these two frequency vectors as inputs of Alice and Bob in the communication game. Since communication $\mathcal{D}^{\to,U}_{1/9}(f^+) \geq \mathcal{D}^{lin,U}_{1/3}(f)$ is required no streaming algorithm with less space exists as otherwise Alice would transfer its state to Bob with less communication. ∎

Using the same proof as in Theorem 33 it follows that all the lower bounds in Section 4 hold for both random streaming models described above.

5.2 Adversarial streams

We now show that any randomized turnstile streaming algorithm for computing $f : \mathbb{F}_2^n \to \mathbb{F}_2$ with error probability δ has to use space that is at least $R^{lin}_{6\delta}(f) - O(\log n + \log(1/\delta))$ under adversarial sequences of updates. The proof is based on the recent line of work that shows that this relationship holds for real-valued sketches [10, 32, 1]. The proof framework developed by [10, 32, 1] for real-valued sketches consists of two steps. First, a turnstile streaming algorithm is converted into a path-independent stream automaton (Definition 35). Second, using the theory of modules and their representations it is shown that such automata can always be represented as linear sketches. We observe that the first step of this framework can be left unchanged under $\mathbb{F}_2$. However, as we show the second step can be significantly simplified as path-independent automata over $\mathbb{F}_2$ can be directly seen as linear sketches without using module theory. Furthermore, since we are working over $\mathbb{F}_2$ we also avoid the $O(\log m)$ factor loss in the reduction between path independent automata and linear sketches that is present in [10].

We use the following abstraction of a *stream automaton* from [10, 32, 1] adapted to our context to represent general turnstile streaming algorithms over $\mathbb{F}_2$.

▶ **Definition 34** (Deterministic Stream Automaton)**.** A *deterministic stream automaton* $\mathcal{A}$ is a Turing machine that uses two tapes, an undirectional read-only input tape and a bidirectional work tape. The input tape contains the input stream σ. After processing the input, the automaton writes an output, denoted as $\phi_{\mathcal{A}}(\sigma)$, on the work tape. A configuration (or state) of $\mathcal{A}$ is determined by the state of its finite control, head position, and contents of the work tape. The computation of $\mathcal{A}$ can be described by a transition function $\oplus_{\mathcal{A}} : C \times \mathbb{F}_2 \to C$, where C is the set of all possible configurations. For a configuration $c \in C$ and a stream σ, we denote by $c \oplus_{\mathcal{A}} \sigma$ the configuration of $\mathcal{A}$ after processing σ starting from the initial configuration c. The set of all configurations of $\mathcal{A}$ that are reachable via processing some input stream σ is denoted as $C(\mathcal{A})$. The space of $\mathcal{A}$ is defined as $\mathcal{S}(\mathcal{A}) = \log |C(\mathcal{A})|$.

We say that a deterministic stream automaton computes a function $f : \mathbb{F}_2^n \to \mathbb{F}_2$ over a distribution Π if $\Pr_{\sigma\sim\Pi}[\phi_{\mathcal{A}}(\sigma) = f(\text{freq } \sigma)] \geq 1 - \delta$.

▶ **Definition 35** (Path-independent automaton)**.** An automaton $\mathcal{A}$ is said to be *path-independent* if for any configuration c and any input stream σ, $c \oplus_{\mathcal{A}} \sigma$ depends only on freq σ and c.

▶ **Definition 36** (Randomized Stream Automaton)**.** A *randomized stream automaton* $\mathcal{A}$ is a deterministic automaton with an additional tape for the random bits. This random tape is initialized with a random bit string R before the automaton is executed. During the execution of the automaton this bit string is used in a bidirectional read-only manner while the rest of the execution is the same as in the deterministic case. A randomized automaton $\mathcal{A}$ is said to be path-independent if for each possible fixing of its randomness R the deterministic automaton $\mathcal{A}_R$ is path-independent. The space complexity of $\mathcal{A}$ is defined as $\mathcal{S}(\mathcal{A}) = \max_R(|R| + \mathcal{S}(\mathcal{A}_R))$.

Theorems 5 and 9 of [32] combined with the observation in Appendix A of [1] that guarantees path independence yields the following:

▶ **Theorem 37** (Theorems 5 and 9 in [32] + [1]). *Suppose that a randomized stream automaton $\mathcal{A}$ computes f on any stream with probability at least $1-\delta$. For an arbitrary distribution Π over streams there exists a deterministic*[15] *path independent stream automaton $\mathcal{B}$ that computes f with probability $1-6\delta$ over Π such that $\mathcal{S}(\mathcal{B}) \le \mathcal{S}(\mathcal{A}) + O(\log n + \log(1/\delta))$.*

The rest of the argument below is based on the work of Ganguly [10] adopted for our needs. Since we are working over a finite field we also avoid the $O(\log m)$ factor loss in the reduction between path independent automata and linear sketches that is present in Ganguly's work.

Let A_n be a path-independent stream automaton over $\mathbb{F}_2$ and let $\oplus$ abbreviate $\oplus_{A_n}$. Define the function $* : \mathbb{F}_2^n \times C(A_n) \to C(A_n)$ as: $x*a = a \oplus \sigma$, where $freq(\sigma) = x$. Let o be the initial configuration of A_n. The *kernel* M_{A_n} of A_n is defined as $M_{A_n} = \{x \in \mathbb{F}_2^n : x * o = 0^n * o\}$.

▶ Proposition 38. The kernel M_{A_n} of a path-independent automaton A_n is a linear subspace of $\mathbb{F}_2^n$.

Proof. For $x, y \in M_{A_n}$ by path independence $(x+y)*o = x*(y*o) = 0^n * o$ so $x+y \in M_{A_n}$. ∎

Since $M_{A_n} \le \mathbb{F}_2^n$ the kernel partitions $\mathbb{F}_2^n$ into cosets of the form $x + M_{A_n}$. Next we show that there is a one to one mapping between these cosets and the states of A_n.

▶ Proposition 39. For $x, y \in \mathbb{F}_2^n$ and a path independent automaton A_n with a kernel M_{A_n} it holds that $x * o = y * o$ if and only if x and y lie in the same coset of M_{A_n}.

Proof. By path independence $x * o = y * o$ iff $x * (x * o) = x * (y * o)$ or equivalently $0^n * o = (x+y) * o$. The latter condition holds iff $x + y \in M_{A_n}$ which is equivalent to x and y lying in the same cost of M_{A_n}. ∎

The same argument implies that the the transition function of a path-independent automaton has to be linear since $(x+y)*o = x*(y*o)$. Combining these facts together we conclude that a path-independent automaton has at least as many states as the best deterministic $\mathbb{F}_2$-sketch for f that succeeds with probability at least $1-6\delta$ over Π (and hence the best randomized sketch as well). Putting things together we get:

▶ **Theorem 40.** *Any randomized streaming algorithm that computes $f : \mathbb{F}_2^n \to \mathbb{F}_2$ under arbitrary updates over $\mathbb{F}_2$ with error probability at least $1-\delta$ has space complexity at least $R_{6\delta}^{lin}(f) - O(\log n + \log(1/\delta))$.*

---

[15] We note that [32] construct $\mathcal{B}$ as a randomized automaton in their Theorem 9 but it can always be made deterministic by fixing the randomness that achieves the smallest error.

Acknowledgements

S. S. thanks Nikhil Mande, Rahul Jain and Anurag Anshu for helpful discussions.

References

1 Yuqing Ai, Wei Hu, Yi Li, and David P. Woodruff. New Characterizations in Turnstile Streams with Applications. In Ran Raz, editor, *31st Conference on Computational Complexity (CCC 2016)*, volume 50 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 20:1–20:22, Dagstuhl, Germany, 2016. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. URL: `http://drops.dagstuhl.de/opus/volltexte/2016/5833`, `doi:http://dx.doi.org/10.4230/LIPIcs.CCC.2016.20`.

2 Noga Alon, Tali Kaufman, Michael Krivelevich, Simon Litsyn, and Dana Ron. Testing reed-muller codes. *IEEE Trans. Information Theory*, 51(11):4032–4039, 2005.

3 Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999. URL: `http://dx.doi.org/10.1006/jcss.1997.1545`, `doi:10.1006/jcss.1997.1545`.

4 Sepehr Assadi, Sanjeev Khanna, and Yang Li. On estimating maximum matching size in graph streams. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19*, pages 1723–1742, 2017.

5 Sepehr Assadi, Sanjeev Khanna, Yang Li, and Grigory Yaroslavtsev. Maximum matchings in dynamic graph streams and the simultaneous communication model. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pages 1345–1364, 2016.

6 Arnab Bhattacharyya, Swastik Kopparty, Grant Schoenebeck, Madhu Sudan, and David Zuckerman. Optimal testing of reed-muller codes. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, pages 488–497, 2010.

7 Eric Blais, Li-Yang Tan, and Andrew Wan. An inequality for the fourier spectrum of parity decision trees. *CoRR*, abs/1506.01055, 2015. URL: `http://arxiv.org/abs/1506.01055`.

8 Andrej Bogdanov and Emanuele Viola. Pseudorandom bits for polynomials. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2007), October 20-23, 2007, Providence, RI, USA, Proceedings*, pages 41–51, 2007.

9 Anirban Dasgupta, Ravi Kumar, and D. Sivakumar. Sparse and lopsided set disjointness via information theory. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 15th International Workshop, APPROX 2012, and 16th International Workshop, RANDOM 2012, Cambridge, MA, USA, August 15-17, 2012. Proceedings*, pages 517–528, 2012.

10 Sumit Ganguly. Lower bounds on frequency estimation of data streams (extended abstract). In *Computer Science - Theory and Applications, Third International Computer Science Symposium in Russia, CSR 2008, Moscow, Russia, June 7-12, 2008, Proceedings*, pages 204–215, 2008.

11 Dmitry Gavinsky, Julia Kempe, and Ronald de Wolf. Quantum communication cannot simulate a public coin. *CoRR*, quant-ph/0411051, 2004. URL: `http://arxiv.org/abs/quant-ph/0411051`.

12 Mohsen Ghaffari and Merav Parter. MST in log-star rounds of congested clique. In *Proceedings of the 2016 ACM Symposium on Principles of Distributed Computing, PODC 2016, Chicago, IL, USA, July 25-28, 2016*, pages 19–28, 2016.

13 Mika Göös and T. S. Jayram. A composition theorem for conical juntas. In *31st Conference on Computational Complexity, CCC 2016, May 29 to June 1, 2016, Tokyo, Japan*, pages 5:1–5:16, 2016.

14 Mika Göös, Shachar Lovett, Raghu Meka, Thomas Watson, and David Zuckerman. Rectangles are nonnegative juntas. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 257–266, 2015.

15 Mika Göös, Toniann Pitassi, and Thomas Watson. Deterministic communication vs. partition number. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 1077–1088, 2015.

16 Parikshit Gopalan, Ryan O'Donnell, Rocco A. Servedio, Amir Shpilka, and Karl Wimmer. Testing fourier dimensionality and sparsity. *SIAM J. Comput.*, 40(4):1075–1100, 2011. URL: `http://dx.doi.org/10.1137/100785429`, `doi:10.1137/100785429`.

17 Vince Grolmusz. On the power of circuits with gates of low {L1} norms. *Theoretical Computer Science*, 188(1–2):117 – 128, 1997. URL: `http://www.sciencedirect.com/science/article/pii/S0304397596002903`, `doi:http://dx.doi.org/10.1016/S0304-3975(96)00290-3`.

18 Hamed Hatami, Kaave Hosseini, and Shachar Lovett. Structure of protocols for XOR functions. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 282–288, 2016.

19 James W. Hegeman, Gopal Pandurangan, Sriram V. Pemmaraju, Vivek B. Sardeshmukh, and Michele Scquizzato. Toward optimal bounds in the congested clique: Graph connectivity and MST. In *Proceedings of the 2015 ACM Symposium on Principles of Distributed Computing, PODC 2015, Donostia-San Sebastián, Spain, July 21 - 23, 2015*, pages 91–100, 2015.

20 Wei Huang, Yaoyun Shi, Shengyu Zhang, and Yufan Zhu. The communication complexity of the hamming distance problem. *Inf. Process. Lett.*, 99(4):149–153, 2006. URL: `http://dx.doi.org/10.1016/j.ipl.2006.01.014`, `doi:10.1016/j.ipl.2006.01.014`.

21 T. S. Jayram. Information complexity: a tutorial. In *Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2010, June 6-11, 2010, Indianapolis, Indiana, USA*, pages 159–168, 2010.

22 T. S. Jayram, Ravi Kumar, and D. Sivakumar. Two applications of information complexity. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing, June 9-11, 2003, San Diego, CA, USA*, pages 673–682, 2003.

23 Tomasz Jurdzinski and Krzysztof Nowicki. MST in $O(1)$ rounds of congested clique. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018*, pages 2620–2632, 2018.

24 Sampath Kannan, Elchanan Mossel, and Grigory Yaroslavtsev. Linear sketching over $\mathbb{F}_2$. *Electronic Colloquium on Computational Complexity (ECCC)*, 23:174, 2016. URL: `http://eccc.hpi-web.de/report/2016/174`.

25 Michael Kapralov, Yin Tat Lee, Cameron Musco, Christopher Musco, and Aaron Sidford. Single pass spectral sparsification in dynamic streams. In *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18-21, 2014*, pages 561–570, 2014.

26 Michael Kapralov, Jelani Nelson, Jakub Pachocki, Zhengyu Wang, David P. Woodruff, and Mobin Yahyazadeh. Optimal lower bounds for universal relation, and for samplers and finding duplicates in streams. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 475–486, 2017.

27 Eyal Kushilevitz and Yishay Mansour. Learning decision trees using the fourier spectrum. *SIAM J. Comput.*, 22(6):1331–1348, 1993. URL: `http://dx.doi.org/10.1137/0222080`, `doi:10.1137/0222080`.

28 Eyal Kushilevitz and Noam Nisan. *Communication complexity*. Cambridge University Press, 1997.

29 Troy Lee and Shengyu Zhang. Composition theorems in communication complexity. In *Automata, Languages and Programming, 37th International Colloquium, ICALP 2010, Bordeaux, France, July 6-10, 2010, Proceedings, Part I*, pages 475–489, 2010.

30 Nikos Leonardos. An improved lower bound for the randomized decision tree complexity of recursive majority,. In *Automata, Languages, and Programming - 40th International Colloquium, ICALP 2013, Riga, Latvia, July 8-12, 2013, Proceedings, Part I*, pages 696–708, 2013.

31 Ming Lam Leung, Yang Li, and Shengyu Zhang. Tight bounds on the randomized communication complexity of symmetric XOR functions in one-way and SMP models. *CoRR*, abs/1101.4555, 2011. URL: `http://arxiv.org/abs/1101.4555`.

32 Yi Li, Huy L. Nguyen, and David P. Woodruff. Turnstile streaming algorithms might as well be linear sketches. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 174–183, 2014.

33 Yang Liu and Shengyu Zhang. Quantum and randomized communication complexity of XOR functions in the SMP model. *Electronic Colloquium on Computational Complexity (ECCC)*, 20:10, 2013. URL: `http://eccc.hpi-web.de/report/2013/010`.

34 Shachar Lovett. Unconditional pseudorandom generators for low degree polynomials. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing, Victoria, British Columbia, Canada, May 17-20, 2008*, pages 557–562, 2008.

35 Shachar Lovett. Recent advances on the log-rank conjecture in communication complexity. *Bulletin of the EATCS*, 112, 2014. URL: `http://eatcs.org/beatcs/index.php/beatcs/article/view/260`.

36 Frédéric Magniez, Ashwin Nayak, Miklos Santha, Jonah Sherman, Gábor Tardos, and David Xiao. Improved bounds for the randomized decision tree complexity of recursive majority. *CoRR*, abs/1309.7565, 2013. URL: `http://arxiv.org/abs/1309.7565`.

37 Frédéric Magniez, Ashwin Nayak, Miklos Santha, and David Xiao. Improved bounds for the randomized decision tree complexity of recursive majority. In *Automata, Languages and Programming - 38th International Colloquium, ICALP 2011, Zurich, Switzerland, July 4-8, 2011, Proceedings, Part I*, pages 317–329, 2011.

38 Andrew McGregor. Graph stream algorithms: a survey. *SIGMOD Record*, 43(1):9–20, 2014. URL: `http://doi.acm.org/10.1145/2627692.2627694`, `doi:10.1145/2627692.2627694`.

39 Ashley Montanaro and Tobias Osborne. On the communication complexity of XOR functions. *CoRR*, abs/0909.3392, 2009. URL: `http://arxiv.org/abs/0909.3392`.

40 Elchanan Mossel, Ryan O'Donnell, and Rocco A. Servedio. Learning juntas. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing, June 9-11, 2003, San Diego, CA, USA*, pages 206–212, 2003.

41 Ryan O'Donnell, John Wright, Yu Zhao, Xiaorui Sun, and Li-Yang Tan. A composition theorem for parity kill number. In *IEEE 29th Conference on Computational Complexity, CCC 2014, Vancouver, BC, Canada, June 11-13, 2014*, pages 144–154, 2014.

42 Ronitt Rubinfeld and Madhu Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM J. Comput.*, 25(2):252–271, 1996.

43 Michael E. Saks and Avi Wigderson. Probabilistic boolean decision trees and the complexity of evaluating game trees. In *27th Annual Symposium on Foundations of Computer Science, Toronto, Canada, 27-29 October 1986*, pages 29–38, 1986.

44 Swagato Sanyal. Near-optimal upper bound on fourier dimension of boolean functions in terms of fourier sparsity. In *Automata, Languages, and Programming - 42nd International Colloquium, ICALP 2015, Kyoto, Japan, July 6-10, 2015, Proceedings, Part I*, pages 1035–1045, 2015.

45 Yaoyun Shi and Zhiqiang Zhang. Communication complexities of symmetric xor functions. *Quantum Inf. Comput*, pages 0808–1762, 2008.

46 Amir Shpilka, Avishay Tal, and Ben lee Volk. On the structure of boolean functions with small spectral norm. In *Innovations in Theoretical Computer Science, ITCS'14, Princeton, NJ, USA, January 12-14, 2014*, pages 37–48, 2014.

47 Xiaoming Sun and Chengu Wang. Randomized communication complexity for linear algebra problems over finite fields. In *29th International Symposium on Theoretical Aspects of Computer Science, STACS 2012, February 29th - March 3rd, 2012, Paris, France*, pages 477–488, 2012.

48 Justin Thaler. Semi-streaming algorithms for annotated graph streams. In *43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy*, pages 59:1–59:14, 2016.

49 Hing Yin Tsang, Chung Hoi Wong, Ning Xie, and Shengyu Zhang. Fourier sparsity, spectral norm, and the log-rank conjecture. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*, pages 658–667, 2013.

50 Emanuele Viola. The sum of d small-bias generators fools polynomials of degree d. In *Proceedings of the 23rd Annual IEEE Conference on Computational Complexity, CCC 2008, 23-26 June 2008, College Park, Maryland, USA*, pages 124–127, 2008.

51 Omri Weinstein and David P. Woodruff. The simultaneous communication of disjointness with applications to data streams. In *Automata, Languages, and Programming - 42nd International Colloquium, ICALP 2015, Kyoto, Japan, July 6-10, 2015, Proceedings, Part I*, pages 1082–1093, 2015.

52 David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014. URL: `http://dx.doi.org/10.1561/0400000060`, `doi:10.1561/0400000060`.

53 Andrew Chi-Chih Yao. Lower bounds by probabilistic arguments (extended abstract). In *24th Annual Symposium on Foundations of Computer Science, Tucson, Arizona, USA, 7-9 November 1983*, pages 420–428, 1983.

54 Grigory Yaroslavtsev. Approximate linear sketching over $\mathbb{F}_2$, 2017.

55 Zhiqiang Zhang and Yaoyun Shi. On the parity complexity measures of boolean functions. *Theor. Comput. Sci.*, 411(26-28):2612–2618, 2010. URL: `http://dx.doi.org/10.1016/j.tcs.2010.03.027`, `doi:10.1016/j.tcs.2010.03.027`.

Appendix

A Information theory

Let X be a random variable supported on a finite set $\{x_1, \ldots, x_s\}$. Let $\mathcal{E}$ be any event in the same probability space. Let $\mathbb{P}[\cdot]$ denote the probability of any event. The *conditional entropy* $H(X \mid \mathcal{E})$ of X conditioned on $\mathcal{E}$ is defined as follows.

▶ **Definition 1** (Conditional entropy).

$$H(X \mid \mathcal{E}) := \sum_{i=1}^{s} \mathbb{P}[X = x_i \mid \mathcal{E}] \log_2 \frac{1}{\mathbb{P}[X = x_i \mid \mathcal{E}]}$$

An important special case is when $\mathcal{E}$ is the entire sample space. In that case the above conditional entropy is referred to as the *Shannon entropy* $H(X)$ of X.

▶ **Definition 2** (Entropy).

$$H(X) := \sum_{i=1}^{s} \mathbb{P}[X = x_i] \log_2 \frac{1}{\mathbb{P}[X = x_i]}$$

Let Y be another random variable in the same probability space as X, taking values from a finite set $\{y_1, \ldots, y_t\}$. Then the conditional entropy of X conditioned on Y, $H(X \mid Y)$, is defined as follows.

▶ **Definition 3.**

$$H(X \mid Y) = \sum_{i=1}^{t} \mathbb{P}[Y = y_i] \cdot H(X \mid Y = y_i)$$

We next define the binary entropy function $H_b(\cdot)$.

▶ **Definition 4** (Binary entropy). For $p \in (0,1)$, the binary entropy of p, $H_b(p)$, is defined to be the Shannon entropy of a random variable taking two distinct values with probabilities p and $1-p$.

$$H_b(p) := p \log_2 \frac{1}{p} + (1-p) \log \frac{1}{1-p}.$$

The following properties of entropy and conditional entropy will be useful.

▶ Fact 5. *(1)* Let X be a random variable supported on a finite set $\mathcal{A}$, and let Y be another random variable in the same probability space. Then $0 \leq H(X \mid Y) \leq H(X) \leq \log_2 |\mathcal{A}|$.

(2) *(Sub-additivity of conditional entropy).* Let $X_1, \ldots, X_n$ be n jointly distributed random variables in some probability space, and let Y be another random variable in the same probability space, all taking values in finite domains. Then,

$$H(X_1, \ldots, X_n \mid Y) \leq \sum_{i=1}^{n} H(X_i \mid Y).$$

(3) Let $X_1, \ldots, X_n$ are independent random variables taking vakues in finite domains. Then,

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i).$$

(4) (Taylor expansion of binary entropy in the neighbourhood of $\frac{1}{2}$).

$$H_b(p) = 1 - \frac{1}{2\log_e 2}\sum_{n=1}^{\infty}\frac{(1-2p)^{2n}}{n(2n-1)}$$

▶ **Definition 6** (Mutual information)**.** Let X and Y be two random variables in the same probability space, taking values from finite sets. The mutual information between X and Y, $I(X;Y)$, is defined as follows.

$$I(X;Y) := H(X) - H(X \mid Y).$$

It can be shown that $I(X;Y)$ is symmetric in X and Y, i.e. $I(X;Y) = I(Y;X) = H(Y) - H(Y \mid X)$.
The following observation follows immediately from the first inequality of Fact 5 (1).

▶ Observation 7. For any two random variables X and Y, $I(X;Y) \le H(X)$.

A.1 Proof of Fact 16

Let $\mathbb{E}X = \delta$. Then,

$$H(X) = \begin{cases} 1 & \text{with probability } \frac{1}{2} + \frac{\delta}{2} \\ -1 & \text{with probability } \frac{1}{2} - \frac{\delta}{2} \end{cases}$$

So,

$$\begin{aligned} H(X) &= H_b\left(\frac{1}{2} + \frac{\delta}{2}\right) \\ &= 1 - \frac{1}{2\log_e 2}\sum_{n=1}^{\infty}\frac{\delta^{2n}}{n(2n-1)} \qquad \text{(From Fact 5 (4))} \\ &\le 1 - \frac{\delta^2}{2}. \end{aligned}$$

B Deterministic $\mathbb{F}_2$-sketching

In the deterministic case it will be convenient to represent $\mathbb{F}_2$-sketch of a function $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$ as a $d \times n$ matrix $M_f \in \mathbb{F}_2^{d\times n}$ that we call the *sketch matrix*. The d rows of M_f correspond to vectors $\alpha_1, \ldots, \alpha_d$ used in the deterministic sketch so that the sketch can be computed as $M_f x$. W.l.o.g below we will assume that the sketch matrix M_f has linearly independent rows and that the number of rows in it is the smallest possible among all sketch matrices (ties in the choice of the sketch matrix are broken arbitrarily).

The following fact is standard (see e.g. [39, 16]):

▶ Fact 8. For any function $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$ it holds that $D^{lin}(f) = dim(f) = rank(M_f)$. Moreover, set of rows of M_f forms a basis for a subspace $A \le \mathbb{F}_2^n$ containing all non-zero coefficients of f.

B.1 Disperser argument

We show that the following basic relationship holds between deterministic linear sketching complexity and the property of being an affine disperser. For randomized $\mathbb{F}_2$-sketching an analogous statement holds for affine extractors as shown in Lemma 16.

▶ **Definition 9** (Affine disperser). A function f is an affine disperser of dimension at least d if for any affine subspace of $\mathbb{F}_2^n$ of dimension at least d the restriction of f on it is a non-constant function.

▶ **Lemma 10.** *Any function $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$ which is an affine disperser of dimension at least d has deterministic linear sketching complexity at least $n-d+1$.*

Proof. Assume for the sake of contradiction that there exists a linear sketch matrix M_f with $k \le n-d$ rows and a deterministic function g such that $g(M_f x) = f(x)$ for every $x \in \mathbb{F}_2^n$. For any vector $b \in \mathbb{F}_2^k$, which is in the span of the columns of M_f, the set of vectors x which satisfy $M_f x = b$ forms an affine subspace of dimension at least $n-k \ge d$. Since f is an affine disperser for dimension at least d the restriction of f on this subspace is non-constant. However, the function $g(M_f x) = g(b)$ is constant on this subspace and thus there exists x such that $g(M_f x) \ne f(x)$, a contradiction. ∎

B.2 Composition and convolution

In order to prove a composition theorem for D^{lin} we introduce the following operation on matrices which for a lack of a better term we call matrix super-slam[16].

▶ **Definition 11** (Matrix super-slam). For two matrices $A \in \mathbb{F}_2^{a\times n}$ and $B \in \mathbb{F}_2^{b\times m}$ their *super-slam* $A \dagger B \in \mathbb{F}_2^{ab^n \times nm}$ is a block matrix consisting of a blocks $(A \dagger B)_i$. The i-th block $(A \dagger B)_i \in \mathbb{F}_2^{b^n \times nm}$ is constructed as follows: for every vector $j \in \{1,\dots,b\}^n$ the corresponding row of $(A \dagger B)_i$ is defined as $(A_{i,1}B_{j_1}, A_{i,2}B_{j_2}, \dots, A_{i,n}B_{j_n})$, where B_k denotes the k^{th} row of B.

▶ Proposition 12. $rank(A \dagger B) \ge rank(A)rank(B)$.

Proof. Consider the matrix C which is a subset of rows of $A \dagger B$ where from each block $(A \dagger B)_i$ we select only b rows corresponding to the vectors j of the form α^n for all $\alpha \in \{1,\dots,b\}$. Note that $C \in \mathbb{F}_2^{ab\times mn}$ and $C_{(i,k),(j,l)} = A_{i,j}B_{k,l}$. Hence, C is a Kronecker product of A and B and we have:

$$rank(A \dagger B) \ge rank(C) = rank(A)rank(B). \qquad ∎$$

The following composition theorem for D^{lin} holds as long as the inner function is balanced:

▶ **Lemma 13.** *For $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$ and $g\colon \mathbb{F}_2^m \to \mathbb{F}_2$ if g is a balanced function then:*

$$D^{lin}(f \circ g) \ge D^{lin}(f)D^{lin}(g)$$

Proof. The multilinear expansions of f and g are given as $f(y) = \sum_{S\in\mathbb{F}_2^n} \hat{f}(S)\chi_S(y)$ and $g(y) = \sum_{S\in\mathbb{F}_2^m} \hat{g}(S)\chi_S(y)$. The multilinear expansion of $f \circ g$ can be obtained as follows. For each monomial $\hat{f}(S)\chi_S(y)$ in the multilinear expansion of f and each variable y_i substitute y_i by the multilinear expansion of g on a set of variables $x_{m(i-1)+1,\dots,mi}$. Multiplying all these multilinear expansions corresponding to the term $\hat{f}(S)\chi_S$ gives a polynomial which is a sum of at most b^n monomials where b is the number of non-zero Fourier coefficients of g. Each such monomial is obtained by picking one monomial from the multilinear expansions corresponding to different variables in χ_S and multiplying them. Note that there are no

[16] This name was suggested by Chris Ramsey.

cancellations between the monomials corresponding to a fixed χ_S. Moreover, since g is balanced and thus $\hat{g}(\emptyset) = 0$ all monomials corresponding to different characters χ_S and $\chi_{S'}$ are unique since S and S' differ on some variable and substitution of g into that variable doesn't have a constant term but introduces new variables. Thus, the characteristic vectors of non-zero Fourier coefficients of $f \circ g$ are the same as the set of rows of the super-slam of the sketch matrices M_f and M_g (note, that in the super-slam some rows can be repeated multiple times but after removing duplicates the set of rows of the super-slam and the set of characteristic vectors of non-zero Fourier coefficients of $f \circ g$ are exactly the same). Using Proposition 12 and Fact 8 we have:

$$D^{lin}(f \circ g) = rank(M_{f \circ g}) = rank(M_f \dagger M_g) \geq rank(M_f) rank(M_g) = D^{lin}(f) D^{lin}(g).$$

∎

Deterministic $\mathbb{F}_2$-sketch complexity of convolution satisfies the following property:

▶ Proposition 14. $D^{lin}(f * g) \leq \min(D^{lin}(f), D^{lin}(g))$.

Proof. The Fourier spectrum of convolution is given as $\widehat{f * g}(S) = \hat{f}(S)\hat{g}(S)$. Hence, the set of non-zero Fourier coefficients of $f * g$ is the intersection of the sets of non-zero coefficients of f and g. Thus by Fact 8 we have $D^{lin}(f * g) \leq \min(rank(M_f, M_g)) = \min(D^{lin}(f), D^{lin}(g))$.
∎

C Randomized $\mathbb{F}_2$-sketching

We represent randomized $\mathbb{F}_2$-sketches as distributions over $d \times n$ matrices over $\mathbb{F}_2$. For a fixed such distribution $\mathcal{M}_f$ the randomized sketch is computed as $\mathcal{M}_f x$. If the set of rows of $\mathcal{M}_f$ satisfies Definition 1 for some reconstruction function g then we call it a *randomized sketch matrix* for f.

C.1 Extractor argument

We now establish a connection between randomized $\mathbb{F}_2$-sketching and affine extractors which will be used to show that the converse of Part 1 of Theorem 14 doesn't hold for arbitrary distributions.

▶ **Definition 15** (Affine extractor). A function $f : \mathbb{F}_2^n \to \mathbb{F}_2$ is an affine δ-extractor if for any affine subspace A of $\mathbb{F}_2^n$ of dimension at least d it satisfies:

$$\min_{z \in \{0,1\}} \Pr_{x \sim U(A)} [f(x) = z] > \delta.$$

▶ **Lemma 16.** *For any $f : \mathbb{F}_2^n \to \mathbb{F}_2$ which is an affine δ-extractor of dimension at least d it holds that:*

$$R_\delta^{lin}(f) \geq n - d + 1.$$

Proof. For the sake of contradiction assume that there exists a randomized linear sketch with a reconstruction function $g : \mathbb{F}_2^k \to \mathbb{F}_2$ and a randomized sketch matrix $\mathcal{M}_f$ which is a distribution over matrices with $k \leq n - d$ rows. First, we show that:

$$\Pr_{x \sim U(\mathbb{F}_2^n) M \sim \mathcal{M}_f} [g(Mx) \neq f(x)] > \delta.$$

Indeed, fix any matrix $M \in supp(\mathcal{M}_f)$. For any affine subspace $\mathcal{S}$ of the form $\mathcal{S} = \{x \in \mathbb{F}_2^n | Mx = b\}$ of dimension at least $n - k \geq d$ we have that $\min_{z \in \{0,1\}} \Pr_{x \sim U(\mathcal{S})}[f(x) = z] > \delta$.

This implies that $\Pr_{x \sim U(\mathcal{S})}[f(x) \neq g(Mx)] > \delta$. Summing over all subspaces corresponding to the fixed M and all possible choices of b we have that $\Pr_{x \sim U(\mathbb{F}_2^n)}[f(x) \neq g(Mx)] > \delta$. Since this holds for any fixed M the bound follows.

Using the above observation it follows by averaging over $x \in \{0,1\}^n$ that there exists $x^* \in \{0,1\}^n$ such that:

$$\Pr_{M \sim \mathcal{M}_f}[g(Mx^*) \neq f(x^*)] > \delta.$$

This contradicts the assumption that $\mathcal{M}_f$ and g form a randomized linear sketch of dimension $k \leq n - d$. ∎

▶ Fact 17. The inner product function $IP(x_1, \dots x_n) = \sum_{i=1}^{n/2} x_{2i-1} \wedge x_{2i}$ is an $(1/2 - \epsilon)$-extractor for affine subspaces of dimension $\geq (1/2 + \alpha)n$ where $\epsilon = \exp(-\alpha n)$.

▶ **Corollary 18.** *Randomized linear sketching complexity of the inner product function is at least $n/2 - O(1)$.*

▶ Remark. We note that the extractor argument of Lemma 16 is often much weaker than the arguments we give in Part 2 and Part 3 Theorem 14 and wouldn't suffice for our applications in Section 4. In fact, the extractor argument is too weak even for the majority function Maj_n. If the first $100\sqrt{n}$ variables of Maj_n are fixed to 0 then the resulting restriction has value 0 with probability $1 - e^{-\Omega(n)}$. Hence for constant error Maj_n isn't an extractor for dimension greater than $100\sqrt{n}$. However, as shown in Section 4.3 for constant error $\mathbb{F}_2$-sketch complexity of Maj_n is linear.

C.2 Existential lower bound for arbitrary distributions

Now we are ready to show that an analog of Part 1 of Theorem 14 doesn't hold for arbitrary distributions, i.e. concentration on a low-dimensional linear subspace doesn't imply existence of randomized linear sketches of small dimension.

▶ **Lemma 19.** *For any fixed constant $\epsilon > 0$ there exists a function $f\colon \mathbb{F}_2^n \to \{+1, -1\}$ such that $R^{lin}_{\epsilon/8}(f) \geq n - 3\log n$ such that f is $(1 - 2\epsilon)$-concentrated on the 0-dimensional linear subspace.*

Proof. The proof is based on probabilistic method. Consider a distribution over functions from $\mathbb{F}_2^n$ to $\{+1, -1\}$ which independently assigns to each x value 1 with probability $1 - \epsilon/4$ and value -1 with probability $\epsilon/4$. By a Chernoff bound with probability $e^{-\Omega(\epsilon 2^n)}$ a random function f drawn from this distribution has at least an $\epsilon/2$-fraction of -1 values and hence $\hat{f}(\emptyset) = \frac{1}{2^n}\sum_{\alpha \in \mathbb{F}_2^n} f(x) \geq 1 - \epsilon$. This implies that $\hat{f}(\emptyset)^2 \geq (1-\epsilon)^2 \geq 1 - 2\epsilon$ so f is $(1-2\epsilon)$-concentrated on a linear subspace of dimension 0. However, as we show below the randomized sketching complexity of some functions in the support of this distribution is large.

The total number of affine subspaces of codimension d is at most $(2 \cdot 2^n)^d = 2^{(n+1)d}$ since each such subspace can be specified by d vectors in $\mathbb{F}_2^n$ and a vector in $\mathbb{F}_2^d$. The number of vectors in each such affine subspace is 2^{n-d}. The probability that less than $\epsilon/8$ fraction of inputs in a fixed subspace have value -1 is by a Chernoff bound at most $e^{-\Omega(\epsilon 2^{n-d})}$. By a union bound the probability that a random function takes value -1 on less than $\epsilon/8$ fraction of the inputs in any affine subspace of codimension d is at most $e^{-\Omega(\epsilon 2^{n-d})} 2^{(n+1)d}$. For $d \leq n - 3\log n$ this probability is less than $e^{-\Omega(\epsilon n)}$. By a union bound, the probability that a random function is either not an $\epsilon/8$-extractor or isn't $(1-2\epsilon)$-concentrated on $\hat{f}(\emptyset)$ is at most $e^{-\Omega(\epsilon n)} + e^{-\Omega(\epsilon 2^n)} \ll 1$. Thus, there exists a function f in the support of our distribution which is an $\epsilon/8$-extractor for any affine subspace of dimension at least $3\log n$ while at the same time is $(1-2\epsilon)$-concentrated on a linear subspace of dimension 0. By Lemma 16 there is no randomized linear sketch of dimension less than $n - 3\log n$ for f which errs with probability less than $\epsilon/8$. ∎

C.3 Random $\mathbb{F}_2$-sketching

The following result is folklore as it corresponds to multiple instances of the communication protocol for the equality function [28, 11] and can be found e.g. in [39] (Proposition 11). We give a proof for completeness.

▶ Fact 20. A function $f : \mathbb{F}_2^n \to \mathbb{F}_2$ such that $\min_{z \in \{0,1\}} \Pr_x[f(x) = z] \le \epsilon$ satisfies

$$R_\delta^{lin}(f) \le \log \frac{\epsilon 2^{n+1}}{\delta}.$$

Proof. We assume that $\mathrm{argmin}_{z \in \{0,1\}} \Pr_x[f(x) = z] = 1$ as the other case is symmetric. Let $T = \{x \in \mathbb{F}_2^n | f(x) = 1\}$. For every two inputs $x \ne x' \in T$ a random $\mathbb{F}_2$-sketch χ_α for $\alpha \sim U(\mathbb{F}_2^n)$ satisfies $\Pr[\chi_\alpha(x) \ne \chi_\alpha(x')] = 1/2$. If we draw t such sketches $\chi_{\alpha_1}, \dots, \chi_{\alpha_t}$ then $\Pr[\chi_{\alpha_i}(x) = \chi_{\alpha_i}(x'), \forall i \in [t]] = 1/2^t$. For any fixed $x \in T$ we have:

$$\Pr[\exists x' \ne x \in T \ \forall i \in [t] : \chi_{\alpha_i}(x) = \chi_{\alpha_i}(x')] \le \frac{|T| - 1}{2^t} \le \frac{\epsilon 2^n}{2^t} \le \frac{\delta}{2}.$$

Conditioned on the negation of the event above for a fixed $x \in T$ the domain of f is partitioned by the linear sketches into affine subspaces such that x is the only element of T in the subspace that contains it. We only need to ensure that we can sketch f on this subspace which we denote as $\mathcal{A}$. On this subspace f is isomorphic to an OR function (up to taking negations of some of the variables) and hence can be sketched using $O(\log 1/\delta)$ uniformly random sketches with probability $1 - \delta/2$. For the OR-function existence of the desired protocol is clear since we just need to verify whether there exists at least one coordinate of the input that is set to 1. In case it does exist a random sketch contains this coordinate with probability $1/2$ and hence evaluates to 1 with probability at least $1/4$. Repeating $O(\log 1/\delta)$ times the desired guarantee follows. ∎

D Towards the proof of Conjecture 3

We call a function $f : \mathbb{F}_2^n \to \{+1, -1\}$ *non-linear* if for all $S \in \mathbb{F}_2^n$ there exists $x \in \mathbb{F}_2^n$ such that $f(x) \ne \chi_S(x)$. Furthermore, we say that f is ϵ-far from being linear if:

$$\max_{S \in \mathbb{F}_2^n} \left[\Pr_{x \sim U(\mathbb{F}_2^n)} [\chi_S(x) = f(x)] \right] = 1 - \epsilon.$$

The following theorem is our first step towards resolving Conjecture 3. Since non-linear functions don't admit 1-bit linear sketches we show that the same is also true for the corresponding communication complexity problem, namely no 1-bit communication protocol for such functions can succeed with a small constant error probability.

▶ **Theorem 21.** *For any non-linear function f that is at most $1/10$-far from linear* $\mathcal{D}^{\rightarrow}_{1/200}(f^+) > 1$.

Proof. Let $S = \arg\max_T \left[\Pr_{x \in \mathbb{F}_2^n}[\chi_T(x) = f(x)\right]$. Pick $z \in \mathbb{F}_2^n$ such that $f(z) \ne \chi_S(z)$. Let the distribution over the inputs (x, y) be as follows: $y \sim U(\mathbb{F}_2^n)$ and $x \sim \mathcal{D}_y$ where D_y is defined as:

$$D_y = \begin{cases} y+z \text{ with probability } 1/2, \\ U(\mathbb{F}_2^n) \text{ with probability } 1/2. \end{cases}$$

Fix any deterministic Boolean function $M(x)$ that is used by Alice to send a one-bit message based on her input. For a fixed Bob's input y he outputs $g_y(M(x))$ for some function g_y that can depend on y. Thus, the error that Bob makes at predicting f for fixed y is at least:

$$\frac{1 - \left|\mathbb{E}_{x\sim D_y}\left[g_y(M(x))f(x+y)\right]\right|}{2}.$$

The key observation is that since Bob only receives a single bit message there are only four possible functions g_y to consider for each y: constants $-1/1$ and $\pm M(x)$.

Bounding error for constant estimators.

For both constant functions we introduce notation $B_y^c = \left|\mathbb{E}_{x\sim D_y}\left[g_y(M(x))f(x+y)\right]\right|$ and have:

$$B_y^c = \left|\mathbb{E}_{x\sim D_y}\left[g_y(M(x))f(x+y)\right]\right| = \left|\mathbb{E}_{x\sim D_y}[f(x+y)]\right| = \left|\frac{1}{2}f(z) + \frac{1}{2}\mathbb{E}_{w\sim U(\mathbb{F}_2^n)}[f(w)]\right|$$

If χ_S is not constant then $\left|\mathbb{E}_{w\sim U(\mathbb{F}_2^n)}[f(w)]\right| \le 2\epsilon$ we have:

$$\left|\frac{1}{2}f(z) + \frac{1}{2}\mathbb{E}_{w\sim U(\mathbb{F}_2^n)}[f(w)]\right| \le \frac{1}{2}\left(|f(z)| + \left|\mathbb{E}_{w\sim U(\mathbb{F}_2^n)}[f(w)]\right|\right) \le 1/2+\epsilon.$$

If χ_S is a constant then w.l.o.g $\chi_S = 1$ and $f(z) = -1$. Also $\mathbb{E}_{w\sim U(\mathbb{F}_2^n)}[f(w)] \ge 1-2\epsilon$. Hence we have:

$$\left|\frac{1}{2}f(z) + \frac{1}{2}\mathbb{E}_{w\sim U(\mathbb{F}_2^n)}[f(w)]\right| = \frac{1}{2}\left|-1 + \mathbb{E}_{w\sim U(\mathbb{F}_2^n)}[f(w)]\right| \le \epsilon.$$

Since $\epsilon \le 1/10$ in both cases $B_y^c \le \frac{1}{2}+\epsilon$ which is the bound we will use below.

Bounding error for message-based estimators.

For functions $\pm M(x)$ we need to bound $\left|\mathbb{E}_{x\sim D_y}\left[M(x)f(x+y)\right]\right|$. We denote this expression as B_y^M. Proposition 22 shows that $\mathbb{E}_y[B_y^M] \le \frac{\sqrt{2}}{2}(1+\epsilon)$.

▶ Proposition 22. $\mathbb{E}_{y\sim U(\mathbb{F}_2^n)}\left[\left|\mathbb{E}_{x\sim D_y}\left[M(x)f(x+y)\right]\right|\right] \le \frac{\sqrt{2}}{2}(1+\epsilon)$.

We have:

$$\begin{aligned}
&\mathbb{E}_y\left[\left|\mathbb{E}_{x\sim D_y}\left[M(x)f(x+y)\right]\right|\right] \\
&= \mathbb{E}_y\left[\left|\frac{1}{2}\left(M(y+z)f(z) + \mathbb{E}_{x\sim D_y}[M(x)f(x+y)]\right)\right|\right] \\
&= \frac{1}{2}\mathbb{E}_y\left[\left|(M(y+z)f(z) + (M*f)(y))\right|\right] \\
&\le \frac{1}{2}\left(\mathbb{E}_y\left[((M(y+z)f(z) + (M*f)(y)))^2\right]\right)^{1/2} \\
&= \frac{1}{2}\left(\mathbb{E}_y\left[((M(y+z)f(z))^2 + ((M*f)(y))^2 + 2M(y+z)f(z)(M*f)(y)))\right]\right)^{1/2} \\
&= \frac{1}{2}\left(\mathbb{E}_y\left[((M(y+z)f(z))^2\right] + \mathbb{E}_y\left[((M*f)(y))^2\right] + \right. \\
&\qquad\qquad \left. 2\mathbb{E}_y\left[M(y+z)f(z)(M*f)(y))\right)\right])^{1/2}
\end{aligned}$$

We have $(M(y+z)f(z))^2 = 1$ and also by Parseval, expression for the Fourier spectrum of convolution and Cauchy-Schwarz:

$$\mathbb{E}_y[((M * f)(y))^2] = \sum_{S \in \mathbb{F}_2^n} \widehat{M * f}(S)^2 = \sum_{S \in \mathbb{F}_2^n} \widehat{M}(S)^2 \hat{f}(S)^2 \leq ||M||_2 ||f||_2 = 1$$

Thus, it suffices to give a bound on $\mathbb{E}[M(y+z)f(z)(M * f)(y))]$. First we give a bound on $(M * f)(y)$:

$$(M * f)(y) = \mathbb{E}_x[M(x)f(x+y)] \leq \mathbb{E}_x[M(x)\chi_S(x+y)] + 2\epsilon$$

Plugging this in we have:

$$\begin{aligned}
&\mathbb{E}_y[M(y+z)f(z)(M * f)(y))] \\
&= -\chi_S(z)\mathbb{E}_y[M(y+z)(M * f)(y))] \\
&\leq -\chi_S(z)\mathbb{E}_y\left[M(y+z)(M * \chi_S)(y)\right] + 2\epsilon \\
&= -\chi_S(z)(M * (M * \chi_S))(z) + 2\epsilon \\
&= -\chi_S(z)^2 \hat{M}(S)^2 + 2\epsilon \\
&\leq 2\epsilon.
\end{aligned}$$

where we used the fact that the Fourier spectrum of $(M * (M * \chi_S))$ is supported on S only and $M * \widehat{(M * \chi_S)}(S) = \hat{M}^2(S)$ and thus $(M * (M * \chi_S))(z) = \hat{M}^2(S)\chi_S(z)$.

Thus, overall, we have:

$$\mathbb{E}_y\left[\left|\mathbb{E}_{x \sim D_y}\left[M(x)f(x+y)\right]\right|\right] \leq \frac{1}{2}\sqrt{2+4\epsilon} \leq \frac{\sqrt{2}}{2}(1+\epsilon). \quad \blacksquare$$

Putting things together.

We have that the error that Bob makes is at least:

$$\mathbb{E}_y\left[\frac{1 - max(B_y^c, B_y^M)}{2}\right] = \frac{1 - \mathbb{E}_y[max(B_y^c, B_y^M)]}{2}$$

Below we now bound $\mathbb{E}_y[max(B_y^c, B_y^M)]$ from above by $99/100$ which shows that the error is at least $1/200$.

$$\begin{aligned}
&\mathbb{E}_y[max(B_y^c, B_y^M)] \\
&= \Pr[B_y^M \geq 1/2+\epsilon]\mathbb{E}[B_y^M | B_y^M \geq 1/2+\epsilon] + Pr[B_y^M < 1/2+\epsilon]\left(\frac{1}{2}+\epsilon\right) \\
&= \mathbb{E}_y[B_y^M] + Pr[B_y^M < 1/2+\epsilon]\left(\frac{1}{2}+\epsilon - \mathbb{E}[B_y^M | B_y^M < 1/2+\epsilon]\right)
\end{aligned}$$

Let $\delta = Pr[B_y^M < 1/2+\epsilon]$. Then the first of the expressions above gives the following bound:

$$\mathbb{E}_y[max(B_y^c, B_y^M)] \leq (1-\delta) + \delta\left(\frac{1}{2}+\epsilon\right) = 1 - \frac{\delta}{2} + \epsilon\delta \leq 1 - \frac{\delta}{2} + \epsilon$$

The second expression gives the following bound:

$$\mathbb{E}_y[max(B_y^c, B_y^M)] \leq \frac{\sqrt{2}}{2}(1+\epsilon) + \delta\left(\frac{1}{2}+\epsilon\right) \leq \frac{\sqrt{2}}{2} + \frac{\delta}{2} + \frac{\sqrt{2}}{2}\epsilon + \epsilon.$$

These two bounds are equal for $\delta = 1 - \frac{\sqrt{2}}{2}(1+\epsilon)$ and hence the best of the two bounds is always at most $(\frac{\sqrt{2}}{4} + \frac{1}{2}) + \epsilon\left(\frac{\sqrt{2}}{4} + 1\right) \leq \frac{99}{100}$ where the last inequality uses the fact that $\epsilon \leq \frac{1}{10}$.

E Auxiliary Proofs

E.1 Proof of Proposition 17

Without loss of generality assume that $p = \Pr[X = 1]$

$$\begin{aligned}
\mathsf{Var}[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\
&= 1 - (\mathbb{E}[X])^2 && (X^2 = 1 \text{ as X is supported on } \{1,\text{-}1\}) \\
&= 1 - (p \cdot 1 + (1-p)(-1))^2 \\
&= 1 - (2p-1)^2) \\
&= 4p(1-p)
\end{aligned}$$

Since $p \leq \frac{1}{2}$, $4(1-p) \in [2,4]$ and the proposition follows.

E.2 Proof of Lemma 21

Let $p \in \mathbb{F}_2[x_1, \ldots, x_n]$ be the $\mathbb{F}_2$-polynomial corresponding to f. Fix one monomial $\mathcal{M} = \Pi_{i \in S} x_i$ of the largest degree. Thus $|S| = d$. We will show that for each assignment $a_{\overline{S}}$ to the variables outside of S, there is an assignment a_S to the variables in S such that $p(a_S, a_{\overline{S}}) = 1$. This will prove that there are at least 2^{n-d} assignments on which p evaluates to 1, and will thus imply the lemma.

To this end, fix an assignment $a_{\overline{S}}$ to the variables in $\overline{S}$. Let $p\,|_{\overline{S} \leftarrow a_{\overline{S}}}$ be the polynomial obtained from p by setting the variables in $\overline{S}$ according to $a_{\overline{S}}$. Notice that since $\mathcal{M}$ was a monomial of largest degree in p, $\mathcal{M}$ continues to be a monomial in $p\,|_{\overline{S} \leftarrow a_{\overline{S}}}$. Thus $p\,|_{\overline{S} \leftarrow a_{\overline{S}}}$ is a non-constant polynomial in the variables $\{x_i \mid i \in S\}$. In particular, this implies that there exists an assignment a_S to the variables in S, such that $p\,|_{\overline{S} \leftarrow a_{\overline{S}}}(a_S) = 1$ (see the discussion in the paragraph after fact 20). This in turn implies that $p(a_S, a_{\overline{S}}) = 1$.