Electronic Colloquium on Computational Complexity, Revision 3 of Report No. 174 (2016)

Linear Sketching over $\mathbb{F}_2$

Sampath Kannan [*] Elchanan Mossel [†] Grigory Yaroslavtsev [‡]

November 6, 2016

Abstract

We initiate a systematic study of linear sketching over $\mathbb{F}_2$. For a given Boolean function $f\colon \{0,1\}^n \to \{0,1\}$ 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. We study a connection between $\mathbb{F}_2$-sketching and a two-player one-way communication game for the corresponding XOR-function. Our results show that this communication game characterizes $\mathbb{F}_2$-sketching under the uniform distribution (up to dependence on error). Implications of this result include: 1) a composition theorem for $\mathbb{F}_2$-sketching complexity of a recursive majority function, 2) a tight relationship between $\mathbb{F}_2$-sketching complexity and Fourier sparsity, 3) lower bounds for a certain subclass of symmetric functions. We also fully resolve a conjecture of Montanaro and Osborne regarding one-way communication complexity of linear threshold functions by designing an $\mathbb{F}_2$-sketch of optimal size.

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 with only a minor loss. 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 doesn't require the stream length to be triply exponential in n and holds for streams of length $\tilde{O}(n)$ constructed through uniformly random updates. Finally, we state a conjecture that asks whether optimal one-way communication protocols for XOR-functions can be constructed as $\mathbb{F}_2$-sketches with only a small loss.

[*] University of Pennsylvania, kannan@cis.upenn.edu

[†] Massachusetts Institute of Technology, elmos@mit.edu

[‡] Indiana University, Bloomington grigory@grigory.us

ISSN 1433-8092

Contents

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 [AMS99]and most recently has been the key to modern randomized algorithms for numerical linear algebra (see survey [Woo14]), graph compression (see survey [McG14]), 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 [HPP$^+$15], 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 [Gan08, LNW14, AHLW16]. 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 [AKLY16], additive norm approximation [AHLW16] and frequency moments [LNW14].

In this paper we study the power of linear sketching over $\mathbb{F}_2$. [1] 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, given a function $f \colon \{0,1\}^n \to \{0,1\}$ that needs to be evaluated over an input $x = (x_1, \dots, x_n)$ we are looking for a distribution over k subsets $\mathbf{S}_1, \dots, \mathbf{S}_k \subseteq [n]$ such that the following holds: for any input x given parities computed over these sets and denoted as $\chi_{\mathbf{S}_1}(x), \chi_{\mathbf{S}_2}(x), \dots, \chi_{\mathbf{S}_k}(x)$[2] it should be possible to compute $f(x)$ with probability $1 - \delta$. In the matrix form sketching corresponds to multiplication over $\mathbb{F}_2$ of the row vector x by a random $n \times k$ matrix whose i-th column is a characteristic vector of the random parity $\chi_{\mathbf{S}_i}$:

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

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 to satisfy the above guarantee.

Definition 1.1 ($\mathbb{F}_2$-sketching)**.** *For a function $f \colon \mathbb{F}_2^n \to \mathbb{F}_2$ we define its* randomized linear sketch complexity[3] *over $\mathbb{F}_2$ with error δ (denoted as $R^{lin}_\delta(f)$) as the smallest integer k such that there*

[1] 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.

[2] Here we use notation $\chi_S(x) = \oplus_{i \in S} x_i$.

[3] 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.

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 $g : \mathbb{F}_2^k \to \mathbb{F}_2$[4] *which satisfies:*

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

As we show in this paper the study of $R_\delta^{lin}(f)$ is closely related to a certain communication complexity 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 complexity problem for XOR-functions.

Definition 1.2 (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*[5] *(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 complexity of XOR-functions has been recently studied extensively in the context of the log-rank conjecture (see e.g. [SZ08, ZS10, MO09, LZ10, LLZ11, SW12, LZ13, TWXZ13, Lov14, HHL16]). 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 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 $i \in [k]$ compute $\chi_{\mathbf{S}_i}(x + y) = \chi_{\mathbf{S}_i}(x) + \chi_{\mathbf{S}_i}(y)$, which is 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 1.3. *Is it true that $R_\delta^{\to}(f^+) = \tilde{\Theta}\left(R_\delta^{lin}(f)\right)$ 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 [DKS12] and a basic upper bound based on random parities [HSZZ06]. Conjecture 1.3 would imply that in order to prove a one-way disjointness lower bound it suffices to only consider $\mathbb{F}_2$-sketches.

In the discussion below using Yao's principle we switch to the equivalent notion of distributional complexity of the above problems denoted as $\mathcal{D}_\delta^{\to}$ and $\mathcal{D}_\delta^{lin}$ respectively. For the formal definitions we refer to the reader to Section 2.1 and a standard textbook on communication complexity [KN97].

[4] If a random family of functions is used here then the definition is changed accordingly. In this paper all g are deterministic.

[5] 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 [KN97] for a formal definition.

Equivalence between randomized and distributional complexities allows us to restate Conjecture 1.3 as $\mathcal{D}^{\to}_{\delta} = \tilde{\Theta}(\mathcal{D}^{lin}_{\delta})$.

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 $\mathcal{D}^{\to,\mu}_{\delta}(f^+)$ (See Section 2 for a formal definition). Our first main result is an analog of Conjecture 1.3 for the uniform distribution U over (x, y) that matches the statement of the conjecture up to dependence on the error probability:

Theorem 1.4. *For any $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$ it holds that $\mathcal{D}^{\to,U}_{\Theta(\frac{1}{n})}(f^+) \ge \mathcal{D}^{lin,U}_{\frac{1}{3}}(f)$.*

A deterministic analog of Definition 1.1 requires that $f(x) = g(\chi_{\alpha_1}(x), \chi_{\alpha_2}(x), \ldots, \chi_{\alpha_k}(x))$ for a fixed choice of $\alpha_1, \ldots, \alpha_k \in \mathbb{F}_2^n$. 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 [MO09] an analog of Conjecture 1.3 holds without any loss in the deterministic case, i.e. $D^{\to}(f^+) = dim(f) = D^{lin}(f)$, where $D^{\to}$ denotes the deterministic one-way communication complexity. This striking fact is one of the reasons why we suggest Conjecture 1.3 as an open problem.

In order to prove Theorem 1.4 we introduce a notion of an *approximate Fourier dimension* (Definition 3.2) 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 3.4 that characterizes both $\mathcal{D}^{lin,U}_{\delta}(f)$ and $\mathcal{D}^{\to,U}_{\delta}(f^+)$ in terms of f's approximate Fourier dimension.

Previous work and our results

Using Theorem 3.4 we derive a number of results that confirm Conjecture 1.3 for specific classes of functions.

Recursive majority For an odd integer n the majority function Maj_n is defined as to be equal 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 [SW86] and serves as an important example of various properties of Boolean functions, most importantly in randomized decision tree complexity ([SW86, JKS03, MNSX11, Leo13, MNS+13]) and most recently deterministic parity decision tree complexity [BTW15].

In Section 4.1 we show to use Theorem 3.4 to obtain the following result:

Theorem 1.5. *For any $\epsilon \in [0, 1]$, $\gamma < \frac{1}{2} - \epsilon$ and $k = \log_3 n$ it holds that:*

$$\mathcal{D}^{\to,U}_{\frac{1}{n}\left(\frac{1}{4}-\epsilon^2\right)}({Maj_3^{\circ k}}^+) \ge \epsilon^2 n + 1.$$

In particular, this confirms Conjecture 1.3 for $Maj_3^{\circ k}$ with at most logarithmic gap as for constant ϵ we get $\mathcal{D}^{\to,U}_{\Theta(\frac{1}{n})}({Maj_3^{\circ k}}^+) = \Omega(n)$. By Yao's principle $R^{\to}_{\Theta(\frac{1}{n})}({Maj_3^{\circ k}}^+) = \Omega(n)$. Using standard error reduction [KN97] for randomized communication this implies that $R_{\delta}^{\to}(Maj_3^{\circ k^+}) = \tilde{\Omega}(n)$ for constant $\delta < 1/2$ almost matching the trivial upper bound.

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 [TWXZ13, STlV14]. A remarkable recent result by Sanyal [San15] 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 1.3 holds (approximately) for 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 $\mathbb{F}_2$.

Applications to streaming In the turnstile streaming model of computation an 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. As we show in Section 5.2 it is easy to see based on the recent work of [Gan08, LNW14, AHLW16] 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 1.4 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 1.3 would imply that space optimal turnstile streaming algorithms over $\mathbb{F}_2$ have to be linear sketches for adversarial streams of length only $2n$.

Linear Threshold Functions Linear threshold functions (LTFs) are one of the most studied classes of Boolean functions as they play a central role in circuit complexity, learning theory and machine learning (See Chapter 5 in [O'D14] for a comprehensive introduction to properties of LTFs). Such functions are parameterized by two parameters θ and m known as threshold and margin respectively (See Definition 6.1 for a formal definition). We design an $\mathbb{F}_2$-sketch for LTFs with complexity $O(\theta/m \log(\theta/m))$. By applying the sketch in the one-way communication setting this fully resolves an open problem posed in [MO09]. Our work shows that dependence on n is not necessary which is an improvement over previously best known protocol due to [LZ13] which achieves communication $O(\theta/m \log n)$. Our communication bound is optimal due to [DKS12]. See Section 6 for details.

Other previous work Closely related to ours is work on communication protocols for XOR-functions started in [SZ08, MO09]. In particular [MO09] presents two basic one-way communication protocols based on random parities. First one, stated as Fact B.7 generalizes the classic protocol for equality. Second one uses the result of Grolmusz [Gro97] 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). 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. However, 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 [MO09, LLZ11, LZ13].

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 [BTW15] and non-adaptive deterministic parity decision trees [STlV14, San15]. 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 [OWZ$^+$14] but the key difference is again adaptivity.

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 1.4. In Section 4 we give applications of this theorem for recursive majority (Theorem 1.5), address function and symmetric functions. In Section 5 we describe applications to streaming. In Section 6 we describe our $\mathbb{F}_2$-sketching protocol for LTFs. In Section 7 we show a lower bound for one-bit protocols making progress towards resolving Conjecture 1.3.

In Appendix A 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 B 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 3.4 doesn't hold for arbitrary distributions. In Appendix C we argue that the parameters of Theorem 3.4 can't be substantially improved.

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 $U(\mathcal{D})$. For a vector x and $p \ge 1$ 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 [Yao83] 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 with it the *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 from $\mathbb{F}_2^n$ to $\mathbb{R}$[6]. 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, $\|f-g\|_2 = \sqrt{\langle f-g, f-g\rangle} = \frac{1}{\sqrt{|\mathbb{F}_2^n|}}\sqrt{\sum_{x\in\mathbb{F}_2^n}(f(x)-g(x))^2}$.

For $x,y\in\mathbb{F}_2^n$ we denote the inner product as $x\cdot y = \sum_{i=1}^n x_iy_i$. 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 $Spec(f) = \{\alpha\in\mathbb{F}_2^n : \hat{f}(\alpha)\neq 0\}$ to denote the set of all non-zero Fourier coefficients of f.

The set of Fourier transforms of functions mapping $\mathbb{F}_2^n\to\mathbb{R}$ forms an inner product space with inner product $\left\langle\hat{f},\hat{g}\right\rangle = \sum_{\alpha\in\mathbb{F}_2^n}\hat{f}(\alpha)\hat{g}(\alpha)$. The corresponding ℓ_2 norm is $\|\hat{f}\|_2 = \sqrt{\left\langle\hat{f},\hat{f}\right\rangle} = \sqrt{\sum_{\alpha\in\mathbb{F}_2^n}\hat{f}(\alpha)^2}$. Note that the inner product and ℓ_2 norm are weighted differently for a function $f:\mathbb{F}_2^n\to\mathbb{R}$ and its Fourier transform $\hat{f}:\mathbb{F}_2^n\to\mathbb{R}$.

Fact 2.1 (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 2.2 (Fourier dimension)**.** *The* Fourier dimension *of $f\colon\mathbb{F}_2^n\to\{+1,-1\}$ denoted as $dim(f)$ is the smallest integer k such that there exists $A\leq\mathbb{F}_2^n$ of dimension k for which $Spec(f)\subseteq A$.*

[6] 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 2.3. *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 2.4. *Fourier coefficients of f^{+z} are given as $\widehat{f^{+z}}(\gamma) = (-1)^{\gamma \cdot z}\hat{f}(\gamma)$ and hence:*

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

Definition 2.5 (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 $f_H^{+z} \colon H \to \mathbb{R}$ for the restriction of f to $H + z$.*

Definition 2.6 (Convolution)**.** *For two functions $f, g \colon \mathbb{F}_2^n \to \mathbb{R}$ their convolution $(f * g) \colon \mathbb{F}_2^n \to \mathbb{R}$ is defined as $(f * g)(x) = \mathbb{E}_{y \sim U(\mathbb{F}_2^n)}[f(x)g(x + y)]$.*

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 [KM93].

Definition 3.1 (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 $\sum_{S \in Z} \hat{f}^2(S) \geq \epsilon$.*

For a function $f \colon \mathbb{F}_2^n \to \{+1, -1\}$ and a parameter $\epsilon > 0$ we introduce a notion of *approximate Fourier dimension* as the smallest integer for which f is ϵ-concentrated on some linear subspace of dimension d.

Definition 3.2 (Approximate Fourier dimension)**.** *Let $\mathcal{A}_k$ be the set of all linear subspaces of $\mathbb{F}_2^n$ of dimension k. For $f \colon \mathbb{F}_2^n \to \{+1, -1\}$ and $\epsilon > 0$ the approximate Fourier dimension $dim_\epsilon(f)$ is defined as:*

$$dim_\epsilon(f) = \min_k \left\{ \exists A \in \mathcal{A}_k \colon \sum_{S \in A} \hat{f}(S)^2 \geq \epsilon \right\}.$$

Definition 3.3 (Approximate Fourier dimension gap)**.** *For $f \colon \mathbb{F}_2^n \to \{+1, -1\}$ and $1 \leq d \leq n$ we define:*

$$\epsilon_d(f) = \max_\epsilon \{dim_\epsilon(f) = d\}, \qquad \Delta_d(f) = \epsilon_d(f) - \epsilon_{d-1}(f),$$

where we refer to $\Delta_d(f)$ as the approximate Fourier dimension gap of dimension d.

The following theorem shows 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 theorem can be applied to all possible values of d and show how to pick specific values of d of interest in Corollary 3.10. We illustrate tightness of Part 3 of this theorem in Appendix C. We also note that the lower bounds given by this theorem are stronger than the basic extractor lower bound given in Appendix B.1. See Remark B.5 for further discussion.

Theorem 3.4. *For any $f \colon \mathbb{F}_2^n \to \{+1,-1\}$, $1 \le d \le n$ and $\epsilon_1 = \epsilon_d(f)$, $\gamma < \frac{1-\sqrt{\epsilon_1}}{2}$, $\delta = \Delta_d(f)/4$:*

1. $\mathcal{D}^{\to,U}_{(1-\epsilon_1)/2}(f^+) \le \mathcal{D}^{lin,U}_{(1-\epsilon_1)/2}(f) \le d,$ 2. $\mathcal{D}^{lin,U}_{\gamma}(f) \ge d+1,$ 3. $\mathcal{D}^{\to,U}_{\delta}(f^+) \ge d.$

Proof. **Part 1[7].** By the assumptions of the theorem we know that there exists a d-dimensional subspace $A \le \mathbb{F}_2^n$ which satisfies $\sum_{S \in A} \hat{f}^2(S) \ge \epsilon_1$. Let $g \colon \mathbb{F}_2^n \to \mathbb{R}$ be a function defined by its Fourier transform as follows:

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

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

Proposition 3.5. *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:

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

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:

$$\begin{aligned} \mathop{\mathbb{E}}_{\theta}\left[\Pr_{x \sim \mathbb{F}_2^n}[g_\theta(x) \ne f(x)]\right] &= \mathop{\mathbb{E}}_{x \sim \mathbb{F}_2^n}\left[\Pr_{\theta}[g_\theta(x) \ne f(x)]\right] \\ &= \mathop{\mathbb{E}}_{x \sim \mathbb{F}_2^n}\left[\Pr_{\theta}[sgn(g(x)-\theta) \ne f(x)]\right] \\ &\le \mathop{\mathbb{E}}_{x \sim \mathbb{F}_2^n}\left[\frac{1}{2}(f(x)-g(x))^2\right] \text{(by Proposition 3.5)} \\ &= \frac{1}{2}\|f-g\|_2^2. \end{aligned}$$

Using the definition of g and Parseval we have:

$$\frac{1}{2}\|f-g\|_2^2 = \frac{1}{2}\|\widehat{f-g}\|_2^2 = \frac{1}{2}\|\hat{f}-\hat{g}\|_2^2 = \frac{1}{2}\sum_{S \notin A} \hat{f}^2(S) \le \frac{1-\epsilon_1}{2}.$$

Thus, there exists a choice of θ such that g_θ achieves error at most $\frac{1-\epsilon_1}{2}$. Clearly g_θ can be computed based on the d parities forming a basis for A and hence $\mathcal{D}^{lin,U}_{(1-\epsilon_1)/2}(f) \le d$.

[7] 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.

Part 2. Fix any deterministic sketch that uses d functions $\chi_{S_1}, \dots, \chi_{S_d}$ and let $S = (S_1, \dots, S_d)$. For fixed values of these sketches $b = (b_1, \dots, b_d)$ where $b_i = \chi_{S_i}(x)$ we denote the restriction on the resulting coset 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:

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

Thus, we have:

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

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

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

The latter sum is the sum of squared Fourier coefficients over a linear subspace of dimension d and hence is at most ϵ_1 by the assumption of the theorem. Using Jensen's inequality:

$$\mathbb{E}_{b\sim U(\mathbb{F}_2^d)}\left[|\widehat{f|_{(S,b)}}(\emptyset)|\right] \leq \sqrt{\mathbb{E}_{b\sim U(\mathbb{F}_2^d)}\left[\widehat{f|_{(S,b)}}(\emptyset)^2\right]} \leq \sqrt{\epsilon_1}.$$

For a fixed restriction (S,b) if $|\hat{f}|_{(S,b)}(\emptyset)| \leq \alpha$ then $|Pr[f|_{(S,b)} = 1] - Pr[f|_{(S,b)} = -1]| \leq \alpha$ and hence no algorithm can predict the value of the restricted function on this coset with probability greater than $\frac{1+\alpha}{2}$. Thus no algorithm can predict $f|_{(S_1,b_1),\dots,(S_d,b_d)}$ for a uniformly random choice of $(b_1, \dots, b_d)$ and hence also on a uniformly at random chosen x with probability greater than $\frac{1+\sqrt{\epsilon_1}}{2}$.

Part 3. Let $\epsilon_2 = \epsilon_{d-1}(f)$ and recall that $\epsilon_1 = \epsilon_d(f)$.

Definition 3.6. *We say that $\mathcal{A} \leq \mathbb{F}_2^n$ distinguishes $x_1, x_2 \in \mathbb{F}_2^n$ if $\exists S \in \mathcal{A}$: $\chi_S(x_1) \neq \chi_S(x_2)$.*

We first prove the following auxiliary lemma.

Lemma 3.7. *Fix $\epsilon_1 > \epsilon_2 \geq 0$ and $x_1, x_2 \in \mathbb{F}_2^n$. If there exists a subspace $\mathcal{A}_d \leq \mathbb{F}_2^n$ of dimension d which distinguishes x_1 and x_2 such that $f : \mathbb{F}_2^n \to \{+1, -1\}$ is ϵ_1-concentrated on $\mathcal{A}_d$ but is not ϵ_2-concentrated on any $d-1$ dimensional linear subspace then:*

$$\Pr_{z\in U(\mathbb{F}_2^n)}[f^{+x_1}(z) \neq f^{+x_2}(z)] \geq \epsilon_1 - \epsilon_2.$$

Proof. Note that for a fixed $x \in \mathbb{F}_2^n$ (by Fact 2.4) the Fourier expansion of f^{+x} can be given as:

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

Thus we have:

$$\begin{aligned}\Pr_{z\in U(\mathbb{F}_2^n)}[f^{+x_1}(z) \neq f^{+x_2}(z)] &= \frac{1}{2}\left(1 - \langle f^{+x_1}, f^{+x_2}\rangle\right) \\ &= \frac{1}{2}\left(1 - \left\langle \sum_{S_1\in\mathbb{F}_2^n} \hat{f}(S_1)\chi_{S_1}\chi_{S_1}(x_1), \sum_{S_2\in\mathbb{F}_2^n} \hat{f}(S_2)\chi_{S_2}\chi_{S_2}(x_2)\right\rangle\right) \\ &= \frac{1}{2}\left(1 - \sum_{S\in\mathbb{F}_2^n} \hat{f}(S)^2\chi_S(x_1)\chi_S(x_2)\right) \text{ (by orthogonality of characters)}\end{aligned}$$

We now analyze the expression $\sum_{S\in\mathbb{F}_2^n} \hat{f}(S)^2\chi_S(x_1)\chi_S(x_2)$. Breaking the sum into two parts we have:

$$\begin{aligned}\sum_{S\in\mathbb{F}_2^n} \hat{f}(S)^2\chi_S(x_1)\chi_S(x_2) &= \sum_{S\in\mathcal{A}_d} \hat{f}(S)^2\chi_S(x_1)\chi_S(x_2) + \sum_{S\notin\mathcal{A}_d} \hat{f}(S)^2\chi_S(x_1)\chi_S(x_2) \\ &\leq \sum_{S\in\mathcal{A}_d} \hat{f}(S)^2\chi_S(x_1)\chi_S(x_2) + (1-\epsilon_1).\end{aligned}$$

To give a bound on the first term we will use the fact that $\mathcal{A}_d$ distinguishes x_1 and x_2. We will need the following simple fact.

Proposition 3.8. *If $\mathcal{A}_d$ distinguishes x_1 and x_2 then there exists a basis $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_d$ in $\mathcal{A}_d$ such that $\chi_{\mathcal{S}_1}(x_1) \neq \chi_{\mathcal{S}_1}(x_2)$ while $\chi_{\mathcal{S}_i}(x_1) = \chi_{\mathcal{S}_i}(x_2)$ for all $i \geq 2$.*

Proof. Since $\mathcal{A}_d$ distinguishes x_1 and x_2 there exists a $\mathcal{S} \in \mathcal{A}_d$ such that $\chi_{\mathcal{S}}(x_1) \neq \chi_{\mathcal{S}}(x_2)$. Fix $\mathcal{S}_1 = \mathcal{S}$ and consider an arbitrary basis in $\mathcal{A}_d$ of the form $(\mathcal{S}_1, \mathcal{T}_2, \dots, \mathcal{T}_d)$. For $i \geq 2$ if $\chi_{\mathcal{T}_i}(x_1) = \chi_{\mathcal{T}_i}(x_2)$ then we let $\mathcal{S}_i = \mathcal{T}_i$. Otherwise, we let $\mathcal{S}_i = \mathcal{T}_i + \mathcal{S}_1$, which preserves the basis and ensures that:

$$\chi_{\mathcal{S}_i}(x_1) = \chi_{\mathcal{T}_i+\mathcal{S}_1}(x_1) = \chi_{\mathcal{T}_i}(x_1)\chi_{\mathcal{S}_1}(x_1) = \chi_{\mathcal{T}_i}(x_1)\chi_{\mathcal{S}_1}(x_2) = \chi_{\mathcal{S}_i}(x_2). \quad \blacksquare$$

Fix the basis $(\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_d)$ in $\mathcal{A}_d$ with the properties given by Proposition 3.8. Let $\mathcal{A}_{d-1} = \mathit{span}(\mathcal{S}_2, \dots, \mathcal{S}_d)$ so that for all $S \in \mathcal{A}_{d-1}$ it holds that $\chi_S(x_1) = \chi_S(x_2)$. Then we have:

$$\begin{aligned}\sum_{S\in\mathcal{A}_d} \hat{f}(S)^2\chi_S(x_1)\chi_S(x_2) &= \sum_{S\in\mathcal{A}_{d-1}} \hat{f}(S)^2\chi_S(x_1)\chi_S(x_2) + \sum_{S\in\mathcal{A}_{d-1}} \hat{f}(S+\mathcal{S}_1)^2\chi_{S+\mathcal{S}_1}(x_1)\chi_{S+\mathcal{S}_1}(x_2) \\ &= \sum_{S\in\mathcal{A}_{d-1}} \hat{f}(S)^2 - \sum_{S\in\mathcal{A}_{d-1}} \hat{f}(S+\mathcal{S}_1)^2\end{aligned}$$

The first term in the above summation is at most ϵ_2 since f is not ϵ_2-concentrated on any $(d-1)$-dimensional linear subspace. The second is at least $\epsilon_1 - \epsilon_2$ since f is ϵ_1-concentrated on $\mathcal{A}_d$.

Thus, putting things together we have that

$$\sum_{S\in\mathbb{F}_2^n} \hat{f}(S)^2\chi_S(x_1)\chi_S(x_2) \leq \epsilon_2 - (\epsilon_1 - \epsilon_2) + (1-\epsilon_1) = 1 - 2(\epsilon_1 - \epsilon_2).$$

This completes that proof showing that $\Pr_{z\in U(\mathbb{F}_2^n)}[f^{+x_1}(z) \neq f^{+x_2}(z)] \geq \epsilon_1 - \epsilon_2$. $\blacksquare$

We are now ready to complete the proof of the third part of Theorem 3.4. We can always assume that the protocol that Alice uses is deterministic since for randomized protocols one can fix their randomness to obtain the deterministic protocol with the smallest error. Fix a $(d-1)$-bit deterministic protocol that Alice is using to send a message to Bob. This protocol partitions the rows of the communication matrix into $t = 2^{d-1}$ rectangles corresponding to different messages. We denote the sizes of these rectangles as $r_1, \ldots, r_t$ and the rectangles themselves as $R_1, \ldots, R_t \subseteq \mathbb{F}_2^n$ respectively. Let the outcome of the protocol be $P(x,y)$. Then the error is given as:

$$\begin{aligned}\mathop{\mathbb{E}}_{x,y\sim U(\mathbb{F}_2^n)}[\mathbb{1}[P(x,y)\neq f(x+y)]] &= \sum_{i=1}^{t}\frac{r_i}{2^n}\times \mathop{\mathbb{E}}_{x\sim U(R_i),y\sim U(\mathbb{F}_2^n)}[\mathbb{1}[P(x,y)\neq f(x+y)]] \\ &\geq \sum_{i\colon r_i>2^{n-d}}\frac{r_i}{2^n}\times \mathop{\mathbb{E}}_{x\sim U(R_i),y\sim U(\mathbb{F}_2^n)}[\mathbb{1}[P(x,y)\neq f(x+y)]],\end{aligned}$$

where we only restricted attention to rectangles of size greater than 2^{n-d}. Our next lemma shows that in such rectangles the protocol makes a significant error:

Lemma 3.9. *If $r_i > 2^{n-d}$ then:*

$$\mathop{\mathbb{E}}_{x\sim U(R_i),y\sim U(\mathbb{F}_2^n)}[\mathbb{1}[P(x,y)\neq f(x+y)]] \geq \frac{1}{2}\frac{r_i-2^{n-d}}{r_i}(\epsilon_1-\epsilon_2).$$

Proof. For $y\in\mathbb{F}_2^n$ let $p_y(R_i)=\min(\Pr_{x\sim U(R_i)}[f(x+y)=1], \Pr_{x\sim U(R_i)}[f(x+y)=-1])$. We have:

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim U(R_i),y\sim U(\mathbb{F}_2^n)}[\mathbb{1}[P(x,y)\neq f(x+y)]] &= \mathop{\mathbb{E}}_{y\sim U(\mathbb{F}_2^n)}\mathop{\mathbb{E}}_{x\sim U(R_i)}[\mathbb{1}[P(x,y)\neq f(x+y)]] \\ &\geq \mathop{\mathbb{E}}_{y\sim U(\mathbb{F}_2^n)}[p_y(R_i)] \\ &\geq \mathop{\mathbb{E}}_{y\sim U(\mathbb{F}_2^n)}[p_y(R_i)(1-p_y(R_i))] \\ &= \mathop{\mathbb{E}}_{y\sim U(\mathbb{F}_2^n)}\left[\frac{1}{2}\Pr_{x_1,x_2\sim U(R_i)}[f(x_1+y)\neq f(x_2+y)]\right] \\ &= \frac{1}{2}\mathop{\mathbb{E}}_{x_1,x_2\sim U(R_i)}\left[\mathop{\mathbb{E}}_{y\sim U(\mathbb{F}_2^n)}[\mathbb{1}\left[f(x_1+y)\neq f(x_2+y)\right]]\right]\end{aligned}$$

Fix a d-dimensional linear subspace $\mathcal{A}_d$ such that g is ϵ_1-concentrated on $\mathcal{A}_d$. There are 2^{n-d} vectors which have the same inner products with all vectors in $\mathcal{A}_d$. Thus with probability at least $\frac{r_i-2^{n-d}}{r_i}$ two random vectors $x_1, x_2 \sim U(R_i)$ are distinguished by $\mathcal{A}_d$. Conditioning on this event we have:

$$\begin{aligned}&\frac{1}{2}\mathop{\mathbb{E}}_{x_1,x_2\sim U(R_i)}\left[\mathop{\mathbb{E}}_{y\sim U(\mathbb{F}_2^n)}[\mathbb{1}\left[f(x_1+y)\neq f(x_2+y)\right]]\right] \\ &\geq \frac{1}{2}\frac{r_i-2^{n-d}}{r_i}\mathop{\mathbb{E}}_{y\sim U(\mathbb{F}_2^n)}[\mathbb{1}\left[f(x_1+y)\neq f(x_2+y)\right]|\mathcal{A}_d \text{ distinguishes } x_1,x_2] \\ &\geq \frac{1}{2}\frac{r_i-2^{n-d}}{r_i}(\epsilon_1-\epsilon_2),\end{aligned}$$

where the last inequality follows by Lemma 3.7. ∎

Using Lemma 3.9 we have:

$$\begin{aligned}
\mathop{\mathbb{E}}_{x,y \sim U(\mathbb{F}_2^n)}\left[\mathbb{1}[P(x,y) \neq f(x+y)]\right] &\geq \frac{\epsilon_1 - \epsilon_2}{2^{n+1}} \sum_{i\colon r_i > 2^{n-d}} \left(r_i - 2^{n-d}\right) \\
&= \frac{\epsilon_1 - \epsilon_2}{2^{n+1}} \left(\sum_{i=1}^{t} \left(r_i - 2^{n-d}\right) - \sum_{i\colon r_i \leq 2^{n-d}} \left(r_i - 2^{n-d}\right) \right) \\
&\geq \frac{\epsilon_1 - \epsilon_2}{2^{n+1}} \left(2^n - 2^{n-1}\right) \\
&= \frac{\epsilon_1 - \epsilon_2}{4},
\end{aligned}$$

where the inequality follows since $\sum_{i=1}^{t} r_i = 2^n$, $t = 2^{d-1}$ and all the terms in the second sum are non-positive. ∎

An important question that arises when applying Theorem 3.4 is the choice of the value of d. The following simple corollaries of Theorem 3.4 give one particularly simple way of choosing these values for any function in such a way that we obtain a non-trivial lower bound for $O(1/n)$-error.

Corollary 3.10. *For any $f\colon \mathbb{F}_2^n \to \{+1,-1\}$ such that $\hat{f}(\emptyset) \leq \theta$ for some constant $\theta < 1$ there exists an integer $d \geq 1$ such that:*

$$\mathcal{D}^{\to,U}_{\Theta(\frac{1}{n})}(f^+) \geq d \geq \mathcal{D}^{lin,U}_{\frac{1}{3}}(f)$$

Proof. We have $\epsilon_0(f) < \theta$ and $\epsilon_n(f) = 1$. Let $d^* = \arg\max_{d=1}^{n} \Delta_d(f)$ and $\Delta(f) = \Delta_{d^*}(f)$. Consider cases:

Case 1. $\Delta(f) \geq \frac{1-\theta}{3}$. By Part 3 of Theorem 3.4 we have that $\mathcal{D}^{\to,U}_{\frac{1-\theta}{12n}}(f^+) \geq d^*$. Furthermore, $\epsilon_{d^*}(f) \geq \theta + \epsilon_{d^*}(f) - \epsilon_{d^*-1}(f) = \theta + \Delta(f) \geq \frac{1}{3} - \frac{2\theta}{3}$. By Part 1 of Theorem 3.4 we have $\mathcal{D}^{lin,U}_{\frac{1-\theta}{3}}(f) \leq d^*$.

Case 2. $\Delta(f) < \frac{1-\theta}{3}$. In this case there exists $d_1 \geq 1$ such that $\epsilon_{d_1}(f) \in [\theta_1, \theta_2]$ where $\theta_1 = \theta + \frac{1-\theta}{3}, \theta_2 = \theta + \frac{2(1-\theta)}{3}$. By averaging there exists $d_2 > d_1$ such that $\Delta_{d_2}(f) = \epsilon_{d_2}(f) - \epsilon_{d_2-1}(f) \geq \frac{1-\theta_2}{n} = \Theta(\frac{1}{n})$. Applying Part 3 of Theorem 3.4 we have that $\mathcal{D}^{\to,U}_{\Theta(\frac{1}{n})}(f^+) \geq d_2$. Furthermore, we have $\epsilon_{d_2}(f) \geq \theta_1$ and hence $\frac{1-\epsilon_{d_2}(f)}{2} \leq \frac{1-\theta_1}{2} < \frac{1-\theta}{3}$. By Part 1 of Theorem 3.4 we have $\mathcal{D}^{lin,U}_{\frac{1-\theta}{3}}(f) \leq d_2$. ∎

The proof of Theorem 1.4 follows directly from Corollary 3.10. If $\theta \leq \frac{1}{3}$ then the statement of the theorem holds. If $\theta \geq \frac{1}{3}$ then $\epsilon_0(f) \geq \frac{1}{3}$ so by Part 1 of Theorem 3.4 we have $\mathcal{D}^{lin,U}_{\frac{1}{3}}(f) \leq 0$ and the inequality holds trivially.

Furthermore, using the same averaging argument as in the proof of Corollary 3.10 we obtain the following generalization of the above corollary that will be useful for our applications.

Corollary 3.11. *For any $f\colon \mathbb{F}_2^n \to \{+1,-1\}$ and d such that $\epsilon_{d-1}(f) \leq \theta$ it holds that:*

$$\mathcal{D}^{\to,U}_{\frac{1-\theta}{4(n-d)}}(f) \geq d.$$

4 Applications

4.1 Composition theorem for majority

In this section using Theorem 3.4 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 A.6) in the randomized case composition results require more work.

Definition 4.1 (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 $Maj_3^{\circ k} \equiv Maj_3 \circ Maj_3 \circ \dots \circ Maj_3$ where the composition is taken k times.

Theorem 4.2. *For any $d \le n$ and $k = \log_3 n$ it holds that $\epsilon_d(Maj_3^{\circ k}) \le \frac{4d}{n}$.*

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

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

Proof. The Fourier expansion of Maj_3 is given as $Maj_3(x_1, x_2, x_3) = \frac{1}{2}(x_1 + x_2 + x_3 - x_1 x_2 x_3)$. For $i \in \{1, 2, 3\}$ let $N_i = \{(i-1)n/3 + 1, \dots, in/3\}$. 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:

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

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

$$\begin{aligned}
&\sum_{Z\in span(S)}\left(\widehat{Maj_3^{\circ k}}(Z)\right)^2 = \frac{1}{4}(\alpha_1+\alpha_2+\alpha_3+\alpha_1\alpha_2\alpha_3) \\
&\leq \frac{1}{4}\left(\alpha_1+\alpha_2+\alpha_3+\frac{1}{3}(\alpha_1+\alpha_2+\alpha_3)\right) = \frac{1}{3}(\alpha_1+\alpha_2+\alpha_3).
\end{aligned}$$

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

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

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

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 4.4.

For a linear subspace $L \leq \mathbb{F}_2^n$ 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 4.4 (Standard subspace domination lemma)**.** *For any linear subspace $L \leq \mathbb{F}_2^n$ of dimension d there exist three standard linear subspaces $S_1, S_2, S_3 \leq \mathbb{F}_2^n$ 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 $k \geq 1$ 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 as 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 can't 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 $\{d+1, \ldots, d+t\}$. 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 as S_3 and clearly $dim(S_3) \leq 2d$. $\blacksquare$

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 4.5. *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 4.2

Proof of Theorem 4.2. By combining Proposition 4.5 and Lemma 4.3 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 4.4) 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 4.2 that proves Theorem 1.5.

Corollary 4.6. *For any $\epsilon \in [0, 1]$, $\gamma < \frac{1}{2} - \epsilon$ and $k = \log_3 n$ it holds that:*

$$\mathcal{D}_\gamma^{lin,U}(Maj_3^{\circ k}) \geq \epsilon^2 n + 1, \qquad \mathcal{D}^{\to,U}_{\frac{1}{n}\left(\frac{1}{4}-\epsilon^2\right)}(Maj_3^{\circ k^+}) \geq \epsilon^2 n + 1.$$

Proof. Fix $d = \epsilon^2 n$. For this choice of d Theorem 4.2 implies that $\epsilon_d(Maj_3^{\circ k}) \leq 4\epsilon^2$. The first part follows from Part 2 of Theorem 3.4. The second part is by Corollary 3.11 as by taking $\epsilon = \sqrt{d/n}$ we can set $\theta = 4\epsilon^2 \geq \epsilon_d(Maj_3^{\circ k})$ and hence:

$$\epsilon^2 n + 1 \leq \mathcal{D}^{\to,U}_{\frac{1-\theta}{4(n-d)}}(Maj_3^{\circ k}) = \mathcal{D}^{\to,U}_{\frac{1-4\epsilon^2}{4n(1-\epsilon^2)}}(Maj_3^{\circ k}) \leq \mathcal{D}^{\to,U}_{\frac{1}{n}\left(\frac{1}{4}-\epsilon^2\right)}(Maj_3^{\circ k}).$$

4.2 Address function and Fourier sparsity

Consider the *addressing function* $Add_n \colon \{0,1\}^{\log n + n} \to \{0,1\}$[8] defined as follows:

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

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. Addressing function has only n^2 non-zero Fourier coefficients. In fact, as shown by Sanyal [San15] Fourier dimension, and hence by Fact A.1 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.

Theorem 4.7. *For the addressing function Add_n and values $1 \le d \le n$ and $\epsilon = d/n$ it holds that:*

$$\mathcal{D}^{lin,U}_{\frac{1-\sqrt{\epsilon}}{2}}(Add_n^+) \ge d, \qquad \mathcal{D}^{\to,U}_{\Theta(\frac{1-\epsilon}{n})}(Add_n) \ge d.$$

Proof. If we apply the standard Fourier notaion 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:

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

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. $\epsilon_d(Add_n) \le \frac{d}{n}$. By Part 2 of Theorem 3.4 and Corollary 3.11 the statement of the theorem follows.

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

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 4.8 (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 $\{\emptyset, \{1, \ldots, n\}\}$:*

$$\mathcal{D}^{lin,U}_{\epsilon/8}(f) \ge \frac{n}{2e}, \qquad \mathcal{D}^{\to,U}_{\Theta(\frac{1-\epsilon}{n})}(f^+) \ge \frac{n}{2e}.$$

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 4.9. *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 4.9 any linear subspace $\mathcal{A}_d$ of dimension at most $d \le n/2$ satisfies that:

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

Thus, f isn't $1 - \epsilon(1 - \frac{ed}{n})$-concentrated on any d-dimensional linear subspace, i.e. $\epsilon_d(f) < 1 - \epsilon(1 - \frac{ed}{n})$. By Part 2 of Theorem 3.4 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 $d \leq \frac{n}{2e}$. The communication complexity lower bound follows by Corollary 3.11 by taking $\theta = \epsilon/8$.

5 Turnstile 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 $\sigma = (\sigma_1, \sigma_2, \dots)$ where $\sigma_i \in \{e_1, \dots, e_n\}$. For a stream σ the resulting vector x corresponds to its frequency vector $\text{freq } \sigma \equiv \sum_i \sigma_i$. Concatenation of two streams σ and τ is denoted as $\sigma \circ \tau$.

5.1 Random streams

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 $\text{freq } \sigma_1 \sim U(\mathbb{F}_2^n)$ and $\text{freq } \sigma_2 \sim U(\mathbb{F}_2^n)$. 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 $\text{freq } \sigma_1$ and $\text{freq } \sigma_2$ are uniformly distributed over $\mathbb{F}_2^n$ and $\sigma = \sigma_1 \circ \sigma_2$.

Theorem 5.1. *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 $f(\text{freq } \sigma)$ with probability at least $1 - \Theta(1/n)$ in the end of the stream has to use space that is at least $\mathcal{D}_{1/3}^{lin,U}(f)$.*

Proof. The proof follows directly from Theorem 1.4 as in both models we can partition the stream into σ_1 and σ_2 such that $\text{freq } \sigma_1$ and $\text{freq } \sigma_2$ 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 $\mathcal{D}_{\Theta(1/n)}^{\to,U}(f^+) \geq \mathcal{D}_{1/3}^{lin,U}(f)$ is required no streaming algorithm with less space exists as otherwise Alice would transfer its state to Bob with less communication. ∎

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 $R_{6\delta}^{lin}(f) - O(\log n + \log(1/\delta))$ 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 [Gan08, LNW14, AHLW16]. The proof framework developed by [Gan08, LNW14, AHLW16] for real-valued sketches consists of two steps. First, a turnstile streaming algorithm is converted into a path-independent stream automaton (Definition 5.3). 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 [Gan08].

We use the following abstraction of a *stream automaton* from [Gan08, LNW14, AHLW16] adapted to our context to represent general turnstile streaming algorithms over $\mathbb{F}_2$.

Definition 5.2 (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 5.3 (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 5.4 (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 [LNW14] combined with the observation in Appendix A of [AHLW16] that guarantees path independence yields the following:

Theorem 5.5 (Theorems 5 and 9 in [LNW14] + [AHLW16])**.** *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*[9] *path independent stream automaton $\mathcal{B}$ that computes f with probability $1 - 6\delta$ over Π such that $\mathcal{S}(\mathcal{B}) \leq \mathcal{S}(\mathcal{A}) + O(\log n + \log(1/\delta))$.*

The rest of the argument below is based on the work of Ganguly [Gan08] 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 $M_{A_n} = \{x \in \mathbb{F}_2^n : x * o = 0^n * o\}$.

[9] We note that [LNW14] 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.

Proposition 5.6. *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 $M_{A_n} \leq \mathbb{F}_2^n$ 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 5.7. *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 5.8. *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 $R_{6\delta}^{lin}(f) - O(\log n + \log(1/\delta))$.*

6 Linear threshold functions

In this section it will be convenient to represent the domain as $\{0, 1\}^n$ rather than $\mathbb{F}_2^n$. We define the sign function *sign*(x) to be 1 if $x \geq 0$ and 0 otherwise.

Definition 6.1. *A monotone linear threshold function (LTF) $f\colon \{0, 1\} \to \{+1, -1\}$ is defined by a collection of weights $w_1 \geq w_2 \cdots \geq w_n \geq 0$ as follows:*

$$f(x_1, \ldots, x_n) = sign\left(\sum_{i=1}^{n} w_i x_i - \theta\right),$$

where θ is called the threshold of the LTF. The *margin of the LTF is defined as:*

$$m = \min_{x \in \{0,1\}^n} \left|\sum_{i=1}^{n} w_i x_i - \theta\right|.$$

W.l.o.g we can assume that LTFs normalized so that $\sum_{i=1}^{n} w_i = 1$. The monotonicity in the above definition is also without loss of generality as for negative weights we can achieve monotonicity by complementing individual bits.

Theorem 6.2. *[MO09] There is a randomized linear sketch for LTFs of size $O((\frac{\theta}{m})^2)$.*

Below we prove the following conjecture.

Conjecture 6.3. *[MO09] There is a randomized linear sketch for LTFs of size $O\left(\frac{\theta}{m} \log\left(\frac{\theta}{m}\right)\right)$.*

In fact, all weights which are below the margin can be completely ignored when evaluating the LTF.

Lemma 6.4. *Let f be a monotone LTF with weights $w_1 \geq w_2 \geq \cdots \geq w_n$, threshold θ and margin m. Let $f^{\geq 2m}$ be an LTF with the same threshold and margin but only restricted to weights $w_1 \geq w_2 \geq \cdots \geq w_t$, where t is the largest integer such that $w_t \geq 2m$. Then $f = f^{\geq m}$.*

Proof. For the sake of contradiction assume there exists an input $(x_1, \ldots, x_n)$ such that $f(x_1, \ldots, x_n) = 1$ while $f^{\geq 2m}(x_1, \ldots, x_t) = 0$. Fix the largest $t^* \geq t$ such that $sign\left(\sum_{i=1}^{t^*} w_i x_i - \theta\right) = 0$ while $sign\left(\sum_{i=1}^{t^*+1} w_i x_i - \theta\right) = 1$. Clearly $w_{t^*+1} \geq 2m$, a contradiction. ∎

The above lemma implies that after dropping the weights which are below $2m$ together with the corresponding variables and reducing the value of n accordingly we can also make the margin equal to $w_n/2$. This observation also gives the following straightforward corollary that proves Conjecture 6.3 about LTFs (up to a logarithmic factor in n).

Corollary 6.5. *There is a randomized linear sketch for LTFs of size $O\left(\frac{\theta}{m} \log n\right)$.*

Proof. We will give a bound on $|\{x \colon f(x) = 0\}|$. If $f(x) = 0$ then $\sum_{i=1}^{n} w_i x_i < \theta$. Since all weights are at least w_n the total number of such inputs is at most $\binom{n}{\theta/w_n} = \binom{n}{\theta/2m} \leq (n+1)^{\theta/2m}$. Thus applying the random $\mathbb{F}_2$-sketching bound (Fact B.7) we get a sketch of size $O\left(\frac{\theta}{m} \log n\right)$ as desired. ∎

Combined with Theorem 6.2 the above corollary proves Conjecture 6.3 except in the case when $\beta \log(\theta/m) < \theta/m < n^{\alpha}$ for all $\alpha > 0$ and $\beta < \infty$. This matches the result of [LZ13].

A full proof of Conjecture 6.3 can be obtained by using hashing to reduce the size of the domain from n down to $poly(\theta/m)$.

Theorem 6.6. *There is a randomized linear sketch for LTFs of size $O\left(\frac{\theta}{m} \log\left(\frac{\theta}{m}\right)\right)$ that succeeds with any constant probability.*

Proof. It suffices to only consider the case when $\theta/m > 100$ since otherwise the bound follows trivially from Theorem 6.2. Consider computing a single linear sketch $\sum_{i \in S} x_i$ where S is a random vector in $\mathbb{F}_2^n$ with each coordinate set to 1 independently with probability $10m^2/\theta^2$. This sketch lets us distinguish the two cases $\|x\|_0 > \theta^2/m^2$ vs. $\|x\|_0 \leq \theta/m$ with constant probability. Indeed:

Case 1. $\|x\|_0 > \theta^2/m^2$. The probability that a set S contains a non-zero coordinate of x in this case is at least:

$$1 - \left(1 - \frac{10m^2}{\theta^2}\right)^{\frac{\theta^2}{m^2}} \geq 1 - (1/e)^{10} > 0.9$$

Conditioned on this event the parity evaluate to 1 with probability at least 1/2. Hence, overall in this case the parity evaluates to 1 with probability at least 0.4.

Case 2. $\|x\|_0 \leq \theta/m$. In this case this probability that S contains a non-zero coordinate and hence the parity can evaluate to 1 is at most:

$$1 - \left(1 - \frac{10m^2}{\theta^2}\right)^{\theta/m} < 1 - (1/2e)^{1/10} < 0.2$$

Thus, a constant number of such sketches allows to distinguish the two cases above with constant probability. If the test above declares that $\|x\|_0 > \theta^2/m^2$ then we output 1 and terminate. Note that conditioned on the test above being correct it never declares that $\|x\|_0 > \theta^2/m^2$ while $\|x\|_0 \leq \theta/m$. Indeed in all such cases, i.e. when $\|x\|_0 > \theta/m$ we can output 1 since if $\|x\|_0 > \theta/m$ then $\sum_{i=1}^n w_i x_i \geq \|x\|_0 w_n \geq \frac{\theta w_n}{m} = 2\theta$, where we used the fact that by Lemma 6.4 we can set $m = w_n/2$.

For the rest of the proof we thus condition on the event that $\|x\|_0 \leq \theta^2/m^2$. By hashing the domain $[n]$ randomly into $O\left(\theta^4/m^4\right)$ buckets we can ensure that no non-zero entries of x collide with any constant probability that is arbitrarily close to 1. This reduces the input length from n down to $O\left(\theta^4/m^4\right)$ and we can apply Corollary 6.5 to complete the proof. [10] ■

This result is also tight as follows from the result of Dasgupta, Kumar and Sivakumar [DKS12] discussed in the introduction. Consider the Hamming weight function $Ham_{\geq d}(x) \equiv \|x\|_0 \geq d$. This function satisfies $\theta = d/n$, $m = 1/2n$. A straightforward reduction from small set disjointness shows that the one-way communication complexity of the XOR-function $Ham_{\geq d}(x \oplus y)$ is $\Omega(d \log d)$. This shows that the bound in Theorem 6.6 can't be improved without any further assumptions about the LTF.

7 Towards the proof of Conjecture 1.3

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 $f(x) \neq \chi_S(x)$. 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 1.3. 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 7.1. *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 $f(z) \neq \chi_S(z)$. 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 $\pm M(x)$.

[10] We note that random hashing doesn't interfere with the linearity of the sketch as it corresponds to treating collections of variables that have the same hash as a single variable representing their sum over $\mathbb{F}_2$. Assuming no collisions this sum evaluates to 1 if and only if a variable of interest is present in the collection.

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 7.2 shows that $\mathbb{E}_y[B_y^M] \le \frac{\sqrt{2}}{2}(1+\epsilon)$.

Proposition 7.2. *$\mathbb{E}_{y \sim U(\mathbb{F}_2^n)}\left[\left|\mathbb{E}_{x \sim D_y}\left[M(x)f(x+y)\right]\right|\right] \le \frac{\sqrt{2}}{2}(1+\epsilon)$.*

We have:

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

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

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

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 $\epsilon \leq \frac{1}{10}$.

References

[AHLW16] 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.

[AKLY16] 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.

[AMS99] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999.

[BTW15] Eric Blais, Li-Yang Tan, and Andrew Wan. An inequality for the fourier spectrum of parity decision trees. *CoRR*, abs/1506.01055, 2015.

[DKS12] 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.

[Gan08] 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.

[GKdW04] Dmitry Gavinsky, Julia Kempe, and Ronald de Wolf. Quantum communication cannot simulate a public coin. *CoRR*, quant-ph/0411051, 2004.

[GOS$^+$11] 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.

[Gro97] Vince Grolmusz. On the power of circuits with gates of low l_1 norms. *Theor. Comput. Sci.*, 188(1-2):117–128, 1997.

[HHL16] Hamed Hatami, Kaave Hosseini, and Shachar Lovett. Structure of protocols for XOR functions. *Electronic Colloquium on Computational Complexity (ECCC)*, 23:44, 2016.

[HPP$^+$15] 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.

[HSZZ06] 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.

[JKS03] 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.

[KM93] Eyal Kushilevitz and Yishay Mansour. Learning decision trees using the fourier spectrum. *SIAM J. Comput.*, 22(6):1331–1348, 1993.

[KN97] Eyal Kushilevitz and Noam Nisan. *Communication complexity.* Cambridge University Press, 1997.

[Leo13] 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.

[LLZ11] 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.

[LNW14] 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.

[Lov14] Shachar Lovett. Recent advances on the log-rank conjecture in communication complexity. *Bulletin of the EATCS*, 112, 2014.

[LZ10] 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.

[LZ13] 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.

[McG14] Andrew McGregor. Graph stream algorithms: a survey. *SIGMOD Record*, 43(1):9–20, 2014.

[MNS$^+$13] 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.

[MNSX11] 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.

[MO09] Ashley Montanaro and Tobias Osborne. On the communication complexity of XOR functions. *CoRR*, abs/0909.3392, 2009.

[O'D14] Ryan O'Donnell. *Analysis of Boolean Functions.* Cambridge University Press, 2014.

[OWZ$^+$14] 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.

[San15] 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.

[STlV14] 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.

[SW86] 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.

[SW12] 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.

[SZ08] Yaoyun Shi and Zhiqiang Zhang. Communication complexities of symmetric xor functions. *Quantum Inf. Comput*, pages 0808–1762, 2008.

[TWXZ13] 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.

[Woo14] David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014.

[Yao83] 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.

[ZS10] Zhiqiang Zhang and Yaoyun Shi. On the parity complexity measures of boolean functions. *Theor. Comput. Sci.*, 411(26-28):2612–2618, 2010.

Appendix

A 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 $\alpha_1, \dots, \alpha_d$ 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. [MO09, GOS$^+$11]):

Fact A.1. *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.*

A.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 B.2.

Definition A.2 (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 A.3. *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. ▮

A.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[11].

Definition A.4 (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 A.5. *$rank(A \dagger B) \ge rank(A)rank(B)$.*

[11] This name was suggested by Chris Ramsey.

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:

$$rank(A \dagger B) \geq rank(C) = rank(A) rank(B). \quad \blacksquare$$

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

Lemma A.6. *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:*

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

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 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 A.5 and Fact A.1 we have:

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

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

Proposition A.7. $D^{lin}(f * g) \leq \min(D^{lin}(f), D^{lin}(g))$.

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 A.1 we have $D^{lin}(f * g) \leq \min(rank(M_f, M_g)) = \min(D^{lin}(f), D^{lin}(g))$. $\blacksquare$

B 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 1.1 for some reconstruction function g then we call it a *randomized sketch matrix* for f.

B.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 3.4 doesn't hold for arbitrary distributions.

Definition B.1 (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 B.2. *For any $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$ 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 B.3. *The inner product function $IP(x_1, \ldots x_n) = \sum_{i=1}^{n/2} x_{2i-1} \wedge x_{2i}$ is an $(1/2 - \epsilon)$-extractor for affine subspaces of dimension $\geq (1/2 + \alpha)n$ where $\epsilon = \exp(-\alpha n)$.*

Corollary B.4. *Randomized linear sketching complexity of the inner product function is at least $n/2 - O(1)$.*

Remark B.5. *We note that the extractor argument of Lemma B.2 is often much weaker than the arguments we give in Part 2 and Part 3 Theorem 3.4 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.*

B.2 Existential lower bound for arbitrary distributions

Now we are ready to show that an analog of Part 1 of Theorem 3.4 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 B.6. *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 B.2 there is no randomized linear sketch of dimension less than $n - 3\log n$ for f which errs with probability less than $\epsilon/8$. ∎

B.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 [KN97, GKdW04] and can be found e.g. in [MO09] (Proposition 11). We give a proof for completeness.

Fact B.7. *A function $f : \mathbb{F}_2^n \to \mathbb{F}_2$ such that $\min_{z \in \{0,1\}} \Pr_x[f(x) = z] \leq \epsilon$ satisfies*

$$R^{lin}_{\delta}(f) \leq \log \frac{\epsilon 2^{n+1}}{\delta}.$$

Proof. We assume that $\operatorname{argmin}_{z \in \{0,1\}} \Pr_x[f(x) = z] = 1$ as the other case is symmetric. Let $T = \{x \in \mathbb{F}_2^n | f(x) = 1\}$. For every two inputs $x \neq x' \in T$ a random $\mathbb{F}_2$-sketch χ_α for $\alpha \sim U(\mathbb{F}_2^n)$ satisfies $\Pr[\chi_\alpha(x) \neq \chi_\alpha(x')] = 1/2$. 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:

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

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. ▮

C Tightness of Theorem 3.4 for the Majority function

An important question is whether Part 3 of Theorem 3.4 is tight. In particular, one might ask whether the dependence on the error probability can be improved by replacing $\Delta_d(f)$ with a larger quantity. As we show below this is not the case and hence Part 3 of Theorem 3.4 is tight.

We consider the majority function Maj_n where n is an odd number. The total Fourier weight on Fourier coefficients corresponding vectors of Hamming weight k is denoted as $W^k(f) = \sum_{\alpha\colon \|\alpha\|_0 = k} \hat{f}(\alpha)^2$. For the majority function it is well-known (see e.g. [O'D14]) that for $\xi = \left(\frac{2}{\pi}\right)^{3/2}$ and odd k it holds that:

$$W^k(Maj_n) = \xi k^{-3/2}(1 \pm O(1/k)).$$

Since Maj_n is a symmetric function whose spectrum decreases monotonically with the Hamming weight of the corresponding Fourier coefficient by a normalization argument as in Lemma 4.9 among all linear subspaces of dimension d the maximum Fourier weight is achieved by the standard subspace $\mathcal{S}_d$ which spans d unit vectors. Computing the Fourier weight of $\mathcal{S}_{n-1}$ we have:

$$\begin{aligned}
\sum_{\alpha \in \mathcal{S}_{n-1}} \widehat{Maj}_n(\alpha)^2 &= 1 - \sum_{\alpha \notin \mathcal{S}_{n-1}} \widehat{Maj}_n(\alpha)^2 \\
&= 1 - \sum_{i=0}^{n/2-1} W^{2i+1}(Maj_n) \frac{\binom{n-1}{2i}}{\binom{n}{2i+1}} \\
&= 1 - \sum_{i=0}^{n/2-1} \xi \frac{1}{(2i+1)^{3/2}} \frac{2i+1}{n} \left(1 \pm O\left(\frac{1}{2i+1}\right)\right) \\
&= 1 - \frac{\gamma}{\sqrt{n}} \pm O\left(\frac{1}{n^{3/2}}\right),
\end{aligned}$$

where $\gamma > 0$ is an absolute constant. Thus, we can set $\epsilon_n(Maj_n) = 1, \epsilon_{n-1}(Maj_n) = 1 - \frac{\gamma}{\sqrt{n}} - O(1/n^{3/2})$ in Part 3 of Theorem 3.4. This gives the following corollary:

Corollary C.1. *It holds that $\mathcal{D}_\delta^{\to,U}(Maj_n^+) \geq n$, where $\delta = \frac{\gamma}{\sqrt{n}} + O\left(\frac{1}{n^{3/2}}\right)$ for some constant $\gamma > 0$.*

Tightness follows from the fact that error $O(1/\sqrt{n})$ for Maj_n can be achieved using a trivial $(n-1)$-bit protocol in which Alice sends the first $n-1$ bits of her input $x_1, \dots, x_{n-1}$ and Bob outputs $Maj_{n-1}(x_1 + y_1, x_2 + y_2, \dots, x_{n-1} + y_{n-1})$. The only inputs on which this protocol can make an error are inputs where there is an equal number of zeros and ones among $x_1 + y_1, \ldots, x_{n-1} + y_{n-1}$. It follows from the standard approximation of binomials that such inputs are an $O(1/\sqrt{n})$ fraction under the uniform distribution.