Efficient Reconstruction of Random Multilinear Formulas

Ankit Gupta * Neeraj Kayal † Satya Lokam ‡

Abstract

In the reconstruction problem for a multivariate polynomial f, we have blackbox access to f and the goal is to efficiently reconstruct a representation of f in a suitable model of computation. We give a polynomial time randomized algorithm for reconstructing *random* multilinear formulas. Our algorithm succeeds with high probability when given blackbox access to the polynomial computed by a random multilinear formula according to a natural distribution. This is the strongest model of computation for which a reconstruction algorithm is presently known, albeit efficient in a distributional sense rather than in the worst-case. Previous results on this problem considered much weaker models such as depth-3 circuits with various restrictions or read-once formulas.

Our proof uses ranks of partial derivative matrices as a key ingredient and combines it with analysis of the algebraic structure of random multilinear formulas. Partial derivative matrices have earlier been used to prove lower bounds in a number of models of arithmetic complexity, including multilinear formulas and constant depth circuits. As such, our results give supporting evidence to the general thesis that mathematical properties that capture efficient computation in a model should also enable learning algorithms for functions efficiently computable in that model.

1 Introduction

We study the problem of reconstructing a multivariate polynomial: given blackbox access to a hidden polynomial $f \in \mathbb{F}[x_1, \ldots, x_n]$ over a finite[1] field $\mathbb{F}$, reconstruct a representation of f in some suitable model of computation. A reconstruction algorithm can adaptively query the blackbox to evaluate f on inputs of its choice from $\mathbb{F}^n$. Its efficiency is measured in terms of the number of queries and the running time. We typically assume f itself to be efficiently computable in some model of computation, e.g., depth-3 circuits of polynomial size, and also require the reconstruction algorithm to produce a succinct representation of f in some (possibly different) model of computation. The most obvious representation of a multivariate polynomial is its formula as a sum, weighted by coefficients from $\mathbb{F}$, of monomials, i.e., a depth-2 $\Sigma\Pi$ formula. In this case, the problem of reconstruction is more commonly referred to as *interpolation*: given blackbox access to a polynomial, produce its representation as a sum of products. However, many interesting polynomials, e.g., determinant, have exponentially long (in the number of variables) representations as a sum of products, whereas as a straight line program or an arithmetic circuit, they can be represented much more succinctly. The reconstruction problem demands such succinct representations as outputs and hence is a generalization of the interpolation problem. In its most general formulation, e.g., produce (roughly) the smallest arithmetic circuit for f, the reconstruction problem is extremely hard. If a circuit class $\mathcal{C}$ has a deterministic reconstruction algorithm, it is easy to see that $\mathcal{C}$ also has a deterministic (blackbox) PIT algorithm. On the other hand, a deterministic PIT implies superpolynomial size lower bounds against $\mathcal{C}$ for an explicit polynomial. Hence, a deterministic reconstruction algorithm for $\mathcal{C}$ is at least as hard as proving superpolynomial lower bounds against $\mathcal{C}$. Thus, much of the research in this area focusses on reconstructing polynomials efficiently computable by weaker variants of arithmetic circuits.

*Microsoft Research India, t-ankitg@microsoft.com
†Microsoft Research India, neeraka@microsoft.com
‡Microsoft Research India, satya@microsoft.com
[1]Many of the definitions make sense for infinite fields as well.

Previous work on the reconstruction problem focussed on polynomials computable by *constant depth* arithmetic circuits and read-once formulas. In particular, depth-2 circuits [KS01], i.e., interpolation problem, depth-3 circuits with bounded top fan-in and multilinear depth-3 formulas with bounded top fan-in [Shp09, KS09]. See [SY10] for more details on previous work.

In this paper, we consider the model of multilinear formulas. An arithmetic formula, using + and $\times$ operations, is *multilinear* if the formal polynomial computed by each of its subformulas is multilinear. Our main result is a randomized reconstruction algorithm for a class of random multilinear formulas. The algorithm uses as a blackbox a multilinear formula randomly chosen according to a natural distribution (see Section 2 below for details). It succeeds with high probability w.r.t. its internal randomness and the choice of the formula from the distribution. Its output is a multilinear formula of the same size as the hidden formula; it is, in fact, the smallest multilinear formula computing the hidden polynomial. This is the strongest model, and the first one of *super-constant depth*, in arithmetic complexity for which an efficient (even in a randomized or distributional sense) reconstruction algorithm is shown. We further remark that a slight variant of the problem of reconstructing multilinear formulas, even for depth three formulas, is known to be NP-hard. Specifically, Hastad [Hås90] showed that reconstructing the smallest set-multilinear formula (an even weaker model than multilinear formulas) for a given set-multilinear polynomial is NP-hard. This indicates that without some kind of a distributional assumption, it would be unrealistic to hope for a reconstruction algorithm for multilinear formulas. Alternatively, it indicates that there is unlikely to be a *worst-case* reconstruction algorithm for multilinear formulas.

From a broad perspective, reconstructing polynomials from arithmetic complexity classes is, in some sense, analogous to learning concept classes of Boolean functions using membership and equivalence queries. (see Chapter 5 of survey by Shpilka and Yehudayaoff [SY10] for justifying arguments for the analogy to the Boolean world and, more generally, for previous work in this area.) While research on the theory of learnability in the Boolean world has evolved into a mature discipline, thanks to fundamental notions such as PAC learning due to Valiant, research on learnability in the arithmetic world has been gaining momentum only in recent years.

A recurring theme in Boolean and arithmetic domains is that techniques used to prove lower bounds for a model of computation are often helpful in designing learning algorithms for that model. At a very high level, a lower bound proof identifies mathematical properties of a model of computation that capture efficient computation in that model. Thus functions efficiently computable in that model should possess the same or similar properties and they should also be useful in learning such functions. This thesis has been borne out in the Boolean world by several examples, e.g., Fourier approximability of AC^0 circuits is useful in both lower bounds and learning algorithms. In the arithmetic world, we see a similar trend, but there are still an abundant number of open questions suggested by this general theme.

Our results in this paper, and in this direction in general, are guided by, and provide supporting evidence to, the thesis mentioned above. One of the key ingredients of our proof is the use of *partial derivative matrices* of polynomials computed in a multilinear formula. We note that properties of partial derivatives of a polynomial have been an important tool in proving lower bounds in a variety of models. In particular, Raz [Raz09] used them to prove lower bounds on multilinear formulas and Raz and Shpilka used them for lower bounds on constant depth circuits. Nisan [Nis91] also used them to prove lower bounds in the noncommutative setting. Thus it is to be expected that properties of partial derivatives of polynomials are useful in reconstruction algorithms. Indeed, Klivans and Shpilka [KS06] prove that whenever the space of partial derivatives has polynomial dimension, one has polynomial time reconstruction algorithms. This implies reconstruction algorithms for some restricted versions of depth-3 circuits and Arithmetic Branching Programs (ABP's) since their partial derivatives span low-dimensional spaces. This approach, however, cannot be used for multilinear formulas since there are multilinear formulas whose partial derivatives span spaces of exponential dimension. Nevertheless, Raz [Raz09] combines rank arguments about partial derivative matrices and combinatorial arguments based on random restrictions to prove quasipolynomial lower bounds on the multilinear formula complexity of the determinant and permanent polynomials. In this paper, too, we exploit rank arguments about partial derivative matrices of polynomials computed

in a multilinear formula and combine them with additional structural properties of random multilinear formulas to derive our reconstruction algorithm.

2 Definitions and Main Result

We recall that an **arithmetic formula** is a binary tree such that (i) each leaf is labeled by either a variable from $X = \{x_1, \ldots, x_n\}$ or an element of the field $\mathbb{F}$, (ii) each internal node is either + gate or $\times$ gate, and (iii) The incoming edges of a + gate are also labeled by constants from $\mathbb{F}$. A + gate computes the linear combination of its inputs with coefficients given by the constants on the incoming edges of the gate. A $\times$ gate computes the product of its inputs. Each gate v in the formula is naturally associated to a polynomial $p_v \in \mathbb{F}[X]$ computed at v. In particular, the polynomial computed at the root (output node) is the polynomial computed by the formula. The *size* of a formula is the number of leaves in the tree. The *(multiplicative) depth* of a node is the number of $\times$ gates on the path from that node to the root. The depth of the formula is the maximum depth of a leaf. An arithmetic formula is said to be **multilinear** if each gate in it computes a multilinear polynomial, i.e., in each of its monomials the power of every input variable is at most one.

Definition 2.1. ***Syntactic Multilinear Formulas:*** *Let Φ be an arithmetic formula over $X = \{x_1, \ldots, x_n\}$. Let Φ_v denote the subformula rooted at a node v and X_v be the set of variables that appear in Φ_v. Then, Φ is said to be* syntactic multilinear *if for every product gate $v = v_1 \times v_2$ of Φ, the sets X_{v_l} and X_{v_2} are disjoint.*

Note that for any multilinear formula, there exists a syntactic multilinear formula of the same size that computes the same polynomial (see [Raz09]). Hence, we often omit the word "syntactic" while referring to multilinear formulas.

A Natural Distribution on the set of Multilinear Formulas:

Our reconstruction algorithm uses, as a blackbox, a random multilinear formula drawn according to a distribution as defined below. Informally, this distribution constructs a binary tree with + and $\times$ gates at alternating levels (with a + gate at the root). Each + gate computes a random linear combination of its inputs over $\mathbb{F}$. Moving down the tree, at each $\times$ gate, we partition the variables into two equal-sized sets and recursively build a subformula rooted at each of this $\times$ gate. We stop the recursion when the number of variables is small enough (we choose this to be about $\log^3 n$ for technical reasons and ensure an error probability of $1/\text{poly}(n)$.)

Note that balanced partitioning of variables at product gates is not a serious loss of generality. This is because if an optimal formula for some polynomial is highly skewed with size s, we can use the depth reduction argument of Valiant et al. for arithmetic circuits and obtain a balanced formula of size at most $s^{O(\log s)}$ and leaves labeled by variables and constants 1 and 0.

A formal definition of the distribution follows:

Let $\mathcal{M}(X, \mathbb{F})$ be the set of all possible syntactic multilinear formulas over the variable set $X = \{x_1, \ldots, x_n\}$ and a (sufficiently large) finite field $\mathbb{F}$. We propose the following method $\texttt{SAMPLE}(X, \mathbb{F})$ to sample a random syntactic multilinear formula from the set $\mathcal{M}(X, \mathbb{F})$, thereby inducing a natural P-samplable distribution $\mathcal{D}(X, \mathbb{F})$ on the set $\mathcal{M}(X, \mathbb{F})$. This distribution also depends on an integer parameter β_n, which we assume to be $\Theta(\log^3 n)$.

Sampling Method $\texttt{SAMPLE}(X, \mathbb{F})$:

Step 1: $\Psi \leftarrow \texttt{CONSTRUCT}(X, +)$, where $\texttt{CONSTRUCT}(X, op)$ is defined below.

Step 2: Let W be the set of wires in Ψ incident to a + gate. Let Φ be the syntactic multilinear arithmetic formula obtained by labeling each $w_i \in W$ by a randomly and independently chosen $c_i \in_R \mathbb{F}$.

Step 3: **return**(Φ).

CONSTRUCT(X, op):

Case 1: $|X| \leq \beta_n$. Let Ψ be the formula with a + gate at the root that has wires incident to it from each $x_i \in X$.

Case 2: $|X| > \beta_n$ and $op = \times$. Partition X randomly into two equal sized sets X_1, X_2 and let $\Psi_1 \leftarrow$ CONSTRUCT($X_1, +$), $\Psi_2 \leftarrow$ CONSTRUCT($X_2, +$). Let Ψ be the formula with a $\times$ gate at the root and Ψ_1, Ψ_2 as its two children.

Case 3: $|X| > \beta_n$ and $op = +$. Let $\Psi_1 \leftarrow$ CONSTRUCT($X, \times$), $\Psi_2 \leftarrow$ CONSTRUCT($X, \times$). Let Ψ be the formula with a + gate at the root and Ψ_1, Ψ_2 as its two children.

Step: **return**(Ψ).

We now state our main reconstruction result for multilinear formulas.

Theorem 2.2. *Let $\Phi \sim \mathcal{D}(X, \mathbb{F})$ be a random multilinear formula sampled as above and let $\hat{\Phi} \in \mathbb{F}[X]$ be the polynomial computed by Φ. Then, there is an $n^{O(1)}$-time randomized algorithm $\mathcal{A}$ which, given blackbox access to $\hat{\Phi}$, constructs a syntactic multilinear formula $\Phi_{\mathcal{A}}$ of size at most size(Φ) and such that*

$$\Pr[\hat{\Phi}_{\mathcal{A}} \neq \hat{\Phi}] \ \leq \ \frac{2^{O(n)}}{|\mathbb{F}|} + \frac{1}{n^{\Omega(1)}},$$

where the probability is taken over the randomness in the choice of Φ and the internal randomness of $\mathcal{A}$.

3 Basic Idea and approach

Suppose we have blackbox access to the output polynomial f of a random multilinear formula Φ. By querying f at points of our choice, we want to recover Φ. How do we do so? We give an overview of our approach to do this.

Determining the nature of the output gate: Let us Observe that if the output node were a $\times$ gate then the output would be a reducible polynomial [2]. The converse is not true in general. That is, it can happen that the output gate is a + gate and f is reducible as well. At this point we invoke the assumption that the formula Φ is chosen randomly and deduce that with high probability over the random choice of Φ the output node is a $\times$ node if and only if f is reducible (Lemma B.3). Thus, we can use the blackbox factoring algorithm of Kaltofen [Kal89] to determine whether f is reducible and this helps us answer our first question. The next thing that we would like to do is get blackbox access to the two children. Once we have that we can recursively do the reconstruction of the two subformulas. There are two cases depending on the nature of the output gate.

Case I: Output node is a $\times$ gate. In this case we factor f using Kaltofen's algorithm. Now it can happen (in rare circumstances) that the number of factors of f is larger than the number of children of the output node. For a generic (i.e. randomly chosen) formula Φ these two quantities will however be equal (Lemma B.3) so that Kaltofen's algorithm provides blackbox access to the two children of the output node. We then recursively compute the formulas for the two children.

Case II: Output node is a + gate. In this case we need to go one level deeper. The two children of the output node are $\times$ gates (except when we are in the base case) so that the output polynomial f is of the form

$$f = A \cdot B + C \cdot D.$$

[2] If one of the children was a constant then the subtree rooted at that node can be discarded and we would have a smaller formula computing the same polynomial

Our aim will be to obtain blackbox access to the four 'grandchildren' A, B, C and D. If we can do that then we can recursively compute formulas for these polynomials and we would be done. At this point we use the fact that we are dealing with (syntactic) multilinear formulas. It means that there exists a partition of the set of variables into four (disjoint) subsets $\bar{u}, \bar{v}, \bar{x}$ and $\bar{y}$ such that

$$f(\bar{u}, \bar{v}, \bar{x}, \bar{y}) = A(\bar{u}, \bar{v}) \cdot B(\bar{x}, \bar{y}) + C(\bar{v}, \bar{x}) \cdot D(\bar{u}, \bar{y}). \tag{1}$$

In general this partition of the set of variables can be arbitrary in which case it becomes much more difficult to find Φ. However, when Φ is random then with high probability all these sets are roughly of the same size (Lemma 5.1). Now it turns out that we can exploit the ideas in the lower bound proof of Raz [] to find this partition of the set of variables. Very roughly, the idea is that for the right partition the rank of a certain related matrix will be very small whereas for every other partition the rank of this matrix will be much larger. This is the one of the key technical arguments (Theorem 5.6) in our work and is described in its proof sketch. For now assume that we know the subsets $\bar{u}, \bar{v}, \bar{x}$ and $\bar{y}$. Knowing these subsets, how do we obtain blackbox access to f? The idea is that if in equation (1) we substitute each $\bar{u}$-variable and each $\bar{v}$-variable to some random values say $\bar{u} = \bar{a}$ and $\bar{v} = \bar{b}$ then $A(\bar{a}, \bar{b})$ becomes a constant so that the degree of $(A \cdot B)$ drops down after this substitution (with high probability, this substitution does not change the degree of $C \cdot D$). This means that the homogeneous part of largest degree of $f(\bar{a}, \bar{b}, \bar{x}, \bar{y})$ is a product of the homogeneous parts of largest degrees of $C(\bar{b}, \bar{x})$ and $D(\bar{a}, \bar{y})$. Thus factoring the homogeneous part of largest degree of f gives us blackbox access to the largest degree homogeneous parts of $C(\bar{b}, \bar{x})$ and $D(\bar{a}, \bar{y})$. This idea can be extended suitably (see Lemma 5.5) to obtain blackbox access to the whole of each polynomial A, B, C and D. This completes our brief overview of the reconstruction algorithm for multilinear formulas.

4 Preliminaries and Notations

Lemma 4.1 (Chernoff's bound)**.** *Let $\zeta_1, \ldots, \zeta_n$ be independent uniform* 0-1 *random variables . Then,*

$$\Pr[\ (1-\delta)n/2 \leq \sum_i \zeta_i \leq (1+\delta)n/2\ ] \geq 1 - 2\exp(-\delta^2 n/8).$$

Lemma 4.2 (DeMillo-Lipton-Schwatz-Zippel)**.** *Let $f \in \mathbb{F}[x_1, \ldots, x_n]$ be a non-zero polynomial of degree $d \geq 0$. Let S be a finite subset of $\mathbb{F}$ and let $r_1, \ldots, r_n$ be selected randomly from S. Then*

$$\Pr[f(r_1, r_2, \ldots, r_n) = 0] \leq \frac{d}{|S|}$$

The above lemma automatically results in the following PIT algorithm which succeeds with probability $\geq 1 - \frac{d}{|S|}$.

Algorithm 1 (Blackbox PIT)**.** *Given blackbox access to a polynomial $f \in \mathbb{F}[x_1, \ldots, x_n]$ of degree d, query $f(r_1, r_2, \ldots, r_n)$ to the blackbox for $r_1, \ldots, r_n \in_R S$, where S is any finite subset of $\mathbb{F}$. Conclude $f = 0$ iff $f(r_1, r_2, \ldots, r_n) = 0$.*

Kaltofen's Blackbox Factoring: We state the multivariate blackbox factoring algorithm by Kaltofen [Kal89] in context of multilinear polynomials,

Lemma 4.3 (Kaltofen's Blackbox Factoring)**.** *There is a randomized polynomial-time algorithm that, given blackbox access to a multilinear polynomial $f \in \mathbb{F}[x_1, \ldots, x_n]$, with probability $1 - 2^{-\Omega(n)}$, outputs blackboxes to all the irreducible factors of f.*

Notation: $[n]$ denotes the set $\{1, 2, \ldots, n\}$. For a polynomial f, $f^{[d]}$ denotes the homogenous degree-d part of f. Tuples would be denoted by placing a bar over a letter, e.g. $\bar{x}$. For a tuple $\bar{\beta} = (\beta_1, \ldots, \beta_n)$, $i\bar{\beta}$ would denote the tuple $(i\beta_1, \ldots, i\beta_n)$. For an arithmetic formula Φ, the polynomial computed at the root is denoted by $\hat{\Phi}$.

5 Reconstructing Multilinear Formulas

5.1 Structural Properties of Multilinear Formulas from $\mathcal{D}(X, \mathbb{F})$

Before we prove Theorem 2.2, we derive some structural properties of random multilinear formulas. Due to space constraints, proofs of the lemmas here appear in Appendix A.

Our first lemma says for the variables in the subformula rooted at a + gate, the two partitions induced by the children ($\times$ gates) of that gate intersect more or less "transversally," i.e., each block of either partition is split nontrivially (in fact in a rather balanced way) by the other partition. Moreover, a child polynomial of a $\times$ gate (a grandchild of the + gate) here is not annihilated by zeroing out either subset of its variables induced by the partition at the sibling product gate.

Lemma 5.1. *Let $\Phi \sim \mathcal{D}(X, \mathbb{F})$. Then, for all nodes of Φ, the following hold with probability at least $1 - \frac{2^{O(n)}}{|\mathbb{F}|} - \frac{1}{n^{\Omega(1)}}$:*

1. *The polynomial computed by a node at (multiplicative) depth h is a homogenous polynomial of degree $\frac{n}{\beta_n 2^h}$.*
2. *The polynomial computed at a + gate is of the form $\alpha.A(\bar{v}, \bar{u})B(\bar{x}, \bar{y}) + \beta.C(\bar{v}, \bar{x})D(\bar{u}, \bar{y})$ where for all $\bar{p} \in \{\bar{v}, \bar{u}, \bar{x}, \bar{y}\}$, $|\bar{p}| \geq \frac{1}{8}|\{\bar{v} \cup \bar{u} \cup \bar{x} \cup \bar{y}\}|$.*
3. *In the above polynomial computed at a + gate, for all $R \in \{A, B, C, D\}$, say $R(\bar{p}, \bar{q})$, $R(\bar{0}, \bar{q}) \neq 0$ and $R(\bar{p}, \bar{0}) \neq 0$.*

Given a multilinear polynomial f over two variable sets $Y = \{y_1, \ldots, y_m\}$ and $Z = \{z_1, \ldots, z_n\}$, define M_f as a $2^m \times 2^n$ matrix whose (p, q) entry, $p \subseteq Y$ and $q \subseteq Z$ is the coefficient of the monomial pq in f. The rank of M_f in this case is denoted by $\text{Rank}_{YZ}(f)$. We will use the following properties of the partial derivatives matrix.

Lemma 5.2 ([Raz09])**.** *Given two multilinear polynomials f and g over the variable set $Y \cup Z$,*

1. $\text{Rank}_{YZ}(f + g) \leq \text{Rank}_{YZ}(f) + \text{Rank}_{YZ}(g)$,
2. $\text{Rank}_{YZ}(f.g) = \text{Rank}_{YZ}(f).\text{Rank}_{YZ}(g)$ *if f and g are polynomials on disjoint sets of variables, and*
3. $\text{Rank}_{YZ}(f) \leq 2^{min(Y(f), Z(f))}$ *where $Y(f)$ and $Z(f)$ are the number of Y and Z variables that occur in f.*

We next show that a random linear combination of two multilinear polynomials can only increase the rank w.h.p.

Lemma 5.3. *Let f and g be two multilinear polynomials over the variable set $Y \cup Z$ and field $\mathbb{F}$. Then for any $S \subset \mathbb{F}$, and two independent random variables α, β,*

$$\Pr_{\alpha, \beta \in_R S}[\text{Rank}_{YZ}(\alpha.f + \beta.g) \geq \max\{\text{Rank}_{YZ}(f), \text{Rank}_{YZ}(g)\}] \geq 1 - \frac{2^{\min\{|Y|, |Z|\}}}{|S|}.$$

5.2 Simulating Blackbox Access to Subformulas

Our reconstruction algorithm will be recursive on the structure of the (unknown, random) multilinear formula. Hence, we will need to simulate blackbox access to its components using blackbox access to the polynomial/formula itself. The next lemma shows this for the homogenous component of a given degree and the theorem below for the grandchildren of a + node.

Lemma 5.4. *Let $\mathbb{F}$ be a field with at least $d+1$ elements and let $f \in \mathbb{F}[x_1, \ldots, x_n]$ be a degree d polynomial. Given blackbox access to f we can simulate blackbox access to $f^{[r]}$'s, where $f^{[r]}$ denotes the homogenous degree-r part of f.*

Proof of this lemma appears in Section A.3.

Theorem 5.5. *Let $\{\{\bar{v}\},\{\bar{u}\},\{\bar{x}\},\{\bar{y}\}\}$ be a partition of $\{x_1,\ldots,x_n\}$ and $f(\bar{v},\bar{u},\bar{x},\bar{y}) = A(\bar{v},\bar{u})B(\bar{x},\bar{y}) + C(\bar{v},\bar{x})D(\bar{u},\bar{y})$ be a non-zero polynomial such that,*

1. *A,B,C,D are homogenous multilinear polynomials over the indicated variable sets,*
2. *either $\deg(AB) \neq \deg(CD)$ or $\deg(A) = \deg(B) = \deg(C) = \deg(D)$,*
3. *for all $R \in \{A,B,C,D\}$, say $R(\bar{p},\bar{q})$, $R(\bar{0},\bar{q}) \neq 0$ and $R(\bar{p},\bar{0}) \neq 0$.*
4. *for all $\bar{p} \in \{\bar{v},\bar{u},\bar{x},\bar{y}\}$, $|\bar{p}| \geq \delta n$, for some $\delta > 0$*

Then there is an $n^{O(1)}$-time randomized algorithm that, given blackbox access to f and the partition $\{\{\bar{v}\},\{\bar{u}\},\{\bar{x}\},\{\bar{y}\}\}$, constructs blackboxes for A,B,C,D with probability at least $1 - \frac{n^{O(1)}}{|\mathbb{F}|} - \frac{1}{2^{\Omega(n)}}$.

Proof Sketch: A detailed proof appears as algorithm `TRICKLEDOWN` in Appendix A.4.

Using Lemma 5.4 and the randomized algorithm for blackbox Polynomial Identity Testing (PIT), we can determine the degrees i for which $f^{[i]} \neq 0$. By (1) and (2), note that there can be at most two such i. Suppose there are two, say i and j. Using PIT, test if $f^{[i]}(\bar{v},\bar{u},\bar{0},\bar{0}) = 0$; if yes, then $f^{[i]} = AB$ and $f^{[j]} = CD$. Otherwise, it is the other way around. Now, we can use Kaltofen's algorithm, and PIT on restrictions of factors of $f^{[i]}$ to determine A and B. For example, if $h(\bar{v},\bar{u},\bar{x},\bar{y})$ is one such factor and $h(\bar{0},\bar{0},\bar{x},\bar{y})$ is 0, then h is a factor of A; else it is a factor of B. We can similarly construct blackboxes for C and D.

Thus the difficult case is when there is a single nonzero $f^{[i]}$ and $\deg(A) = \deg(B) = \deg(C) = \deg(D) =: d$. Note that $f(\bar{v},\bar{u},\bar{0},\bar{0}) = C(\bar{v},\bar{0})D(\bar{u},\bar{0})$. It follows that using Kaltofen and PIT as before, we can construct blackboxes for $C(\bar{v},\bar{0})$ and $D(\bar{u},\bar{0})$ (but not for full C and D). Similarly, we can construct blackboxes for $C(\bar{0},\bar{x})$ and $D(\bar{0},\bar{y})$. We can also immediately determine the degree d as $d = \deg(C) = \log\left(\frac{C(2\bar{\alpha},\bar{0})}{C(\bar{\alpha},\bar{0})}\right)$ for a randomly chosen $\bar{\alpha} \in \mathbb{F}^{|\bar{v}|}$.

Suppose now, we want to determine $C(\bar{\alpha},\bar{\beta})$ for $\bar{\alpha} \in \mathbb{F}^{|\bar{v}|}$, $\bar{\beta} \in \mathbb{F}^{|\bar{x}|}$. Choose a random $\bar{\gamma} \in_R \mathbb{F}^{|\bar{y}|}$ and for $g \in \{A,B,C,D,f\}$, denote by $\hat{g}$ the restriction of g by fixing $\bar{x}$ to $\bar{\beta}$ and $\bar{y}$ to $\bar{\gamma}$. Then, we can see that $\hat{f}(\bar{v},\bar{u})^{[2d]} = \hat{C}^{[d]}(\bar{v})\hat{D}^{[d]}(\bar{u})$ (since $\hat{B}$ becomes a constant and $\hat{A}\hat{B}$ contributes only to lower degree terms of $\hat{f}$). Using Kaltofen and PIT and blackbox for $\hat{f}^{[2d]}$, we can construct blackboxes for $\hat{C}^{[d]}(\bar{v})$ and $\hat{D}^{[d]}(\bar{u})$. Note that we want $\hat{C}(\bar{\alpha})$ and that

$$\hat{C}(\bar{v}) = \hat{C}^{[d]}(\bar{v}) + \ldots + \hat{C}^{[1]}(\bar{v}) + C(\bar{0},\bar{\beta}) \quad \text{and} \quad \hat{D}(\bar{u}) = \hat{D}^{[d]}(\bar{u}) + \ldots + \hat{D}^{[1]}(\bar{u}) + D(\bar{0},\bar{\gamma}). \tag{2}$$

Recall that we already have blackboxes for $C(\bar{0},\bar{x})$ and $D(\bar{0},\bar{y})$. We now need to get blackboxes $\hat{C}^{[d-1]},\ldots,\hat{C}^{[1]}$ and similarly for $\hat{D}$. To this end, consider the following equations:

$$\begin{aligned} \hat{f}(\bar{v},\bar{u})^{[2d-1]} &= \hat{C}^{[d]}(\bar{v})\hat{D}^{[d-1]}(\bar{u}) + \hat{C}^{[d-1]}(\bar{v})\hat{D}^{[d]}(\bar{u}), \\ \hat{f}(\bar{v},2\bar{u})^{[2d-1]} &= 2^{d-1}\hat{C}^{[d]}(\bar{v})\hat{D}^{[d-1]}(\bar{u}) + 2^{d}\hat{C}^{[d-1]}(\bar{v})\hat{D}^{[d]}(\bar{u}). \end{aligned}$$

Since we already have blackboxes for $\hat{C}^{[d]}$, $\hat{D}^{[d]}$, and $\hat{f}^{[2d-1]}$, we can solve these equations to get blackboxes for $\hat{C}^{[d-1]}$ (by setting $\bar{u}$ randomly) and $\hat{D}^{[d-1]}$ (by setting $\bar{v}$ randomly). Using similar, but somewhat more involved, equations, we obtain blackboxes for $\hat{C}^{[i]}(\bar{v})$ and $\hat{D}^{[i]}(\bar{v})$ for $1 \leq i \leq d-2$ (see Appendix A.4 for details). Using these in equation (2), we get blackbox for $\hat{C}(\bar{v})$ and hence can evaluate $\hat{C}(\bar{\alpha})$.

A similar argument can be used to gain blackbox access A, B, and D. ■

5.3 The Reconstruction Algorithm RECONSTRUCT$(\mathcal{O}_{\hat{\Phi}}, X, \mathbb{F}, m)$

We are now ready to present the reconstruction algorithm for random multilinear formulas.
Input: oracle $\mathcal{O}_{\hat{\Phi}}$ for polynomial $\hat{\Phi}$ computable by a multilinear formula Φ sampled using SAMPLE$(X, \mathbb{F})$ where $X = \{x_1, \ldots, x_n\}$ and size m of the seed partition[3]($m = \Theta(\log n)$).
Output: multilinear formula Ψ such that $|\Psi| \leq |\Phi|$ and $\hat{\Psi} = \hat{\Phi}$, or else FAIL.

Step 1 *Determining linearity*: For any $x_i \in X$, $f_i = \hat{\Phi}|_{x_i=1} - \hat{\Phi}|_{x_i=0}$ is the coefficient polynomial of x_i in $\hat{\Phi}$. For all f_i's, using blackbox PIT on $f_i|_{x_j=1} - f_i|_{x_j=0}$, determine if f_i depends on x_j. If for all x_i with a non-zero f_i, f_i does not depend on X, then $\hat{\Phi}$ is linear and in this case simply interpolate $\hat{\Phi}$ exactly and output a Σ-circuit for it.

Step 2 *Reducible* $\hat{\Phi}$: Using Kaltofen's factoring algorithm construct oracles for irreducible factors h_i's, of $\hat{\Phi}$. If $\hat{\Phi}$ is irreducible proceed to the next step. Else using blackbox PIT, as described in the previous step, determine the variable sets of these factors. Recursively using RECONSTRUCT, construct formulas Ψ_i's for h_i's. If RECONSTRUCT fails on any h_i output FAIL. Else, output a formula with $\times$ gate at the root and Ψ_i's as its children.

Step 3 *Determining a seed partition*: Let $\hat{\Phi} = A(\bar{v}, \bar{u})B(\bar{x}, \bar{y}) + C(\bar{v}, \bar{x})D(\bar{u}, \bar{y})$. Randomly choose an m-sized subset S of X. In $\hat{\Phi}$, instantiate the variables in $X \setminus S$ to random values over $\mathbb{F}$ to get $\hat{\Phi}_S = A_S(\bar{v}_S, \bar{u}_S)B_S(\bar{x}_S, \bar{y}_S) + C_S(\bar{v}_S, \bar{x}_S)D_S(\bar{u}_S, \bar{y}_S)$ and interpolate it in $n^{O(1)}$ time. Iterate over all possible partitions $\{\{\bar{v}''\}, \{\bar{u}''\}, \{\bar{x}''\}, \{\bar{y}''\}\}$ of S such that the size of each set in them is at least γm (for a small enough γ) and let $\{\{\bar{v}'\}, \{\bar{u}'\}, \{\bar{x}'\}, \{\bar{y}'\}\}$ be a partition such that $\text{Rank}_{\{\bar{v}'\}\{\bar{y}'\}}(\hat{\Phi}_S|_{\bar{v}',\bar{y}'}) \leq 2$ and $\text{Rank}_{\{\bar{u}'\}\{\bar{x}'\}}(\hat{\Phi}_S|_{\bar{u}',\bar{x}'}) \leq 2$ where $\hat{\Phi}_S|_{\bar{v}',\bar{y}'}$ is $\hat{\Phi}_S$ with variables in $S \setminus \{\bar{v}', \bar{y}'\}$ instantiated to random values in $\mathbb{F}$ and similarly for $\hat{\Phi}_S|_{\bar{u}',\bar{x}'}$. This can be done in $n^{O(1)}$ time, having interpolated $\hat{\Phi}_S$, as there are $2^{O(\log n)}$ such possible partitions and the partial derivative matrix on $O(\log n)$ variables is of size at most $2^{O(\log n)}$.

Step 4 *Extending the seed partition* $\{\{\bar{v}'\}, \{\bar{u}'\}, \{\bar{x}'\}, \{\bar{y}'\}\}$: For all $x_i \in X \setminus S$ do the following. Let $S_i = S \cup \{x_i\}$. In $\hat{\Phi}$, instantiate the variables in $X \setminus S_i$ to random values over $\mathbb{F}$ to get $\hat{\Phi}_{S_i}$ and interpolate it in $2^{O(\log n)}$ time. Iterate over the following 4 partitions of S_i, $\{\{\bar{v}', x_i\}, \{\bar{u}'\}, \{\bar{x}'\}, \{\bar{y}'\}\}$, $\{\{\bar{v}'\}, \{\bar{u}', x_i\}, \{\bar{x}'\}, \{\bar{y}'\}\}$, $\{\{\bar{v}'\}, \{\bar{u}'\}, \{\bar{x}', x_i\}, \{\bar{y}'\}\}$, $\{\{\bar{v}'\}, \{\bar{u}'\}, \{\bar{x}'\}, \{\bar{y}', x_i\}\}$ and determine the partition $\{\{\bar{v}''\}, \{\bar{u}''\}, \{\bar{x}''\}, \{\bar{y}''\}\}$ such that,
$\text{Rank}_{\{\bar{v}''\}\{\bar{y}''\}}(\hat{\Phi}_{S_i}|_{\bar{v}'',\bar{y}''}) \leq 2$ and $\text{Rank}_{\{\bar{u}''\}\{\bar{x}''\}}(\hat{\Phi}_{S_i}|_{\bar{u}'',\bar{x}''}) \leq 2$ where $\hat{\Phi}_{S_i}|_{\bar{v}'',\bar{y}''}$ is $\hat{\Phi}_{S_i}$ with variables in $S_i \setminus \{\bar{v}'', \bar{y}''\}$ instantiated to random values in $\mathbb{F}$. Attach x_i to the appropriate block of the seed partition. This can be done in $2^{O(\log n)}$ time.

Step 5: Using TRICKLEDOWN algorithm and the above determined partition $\{\{\bar{v}\}, \{\bar{u}\}, \{\bar{x}\}, \{\bar{y}\}\}$ of X construct oracles for A,B,C,D. Then, recursively using RECONSTRUCT, construct formulas Ψ_R's for $R \in \{A, B, C, D\}$. If RECONSTRUCT fails on for any of them output FAIL. Else, let Ψ_{AB} be the formula with $\times$ gate at the root and Ψ_A, Ψ_B as its children. Output a formula Ψ with $+$ gate at the root and Ψ_{AB}, Ψ_{CD} as its children.

This completes the description of the algorithm RECONSTRUCT. Algorithm $\mathcal{A}$ of Theorem 2.2 is now essentially RECONSTRUCT, returning Ψ using blackbox calls to $\hat{\Phi}$. (If RECONSTRUCT outputs FAIL, $\mathcal{A}$ outputs a random multilinear formula.) The bound on the running time of $\mathcal{A}$ is obvious. For correctness, it's crucial to show that the partition determined by steps 3 and 4 is, w.h.p., the original partition of Φ. We do this in the next section. This will complete the proof of Theorem 2.2. ■

[3]size of the seed partition is kept unchanged while recursing.

5.4 Uniqueness of the Seed Partition

In this section, we discuss Steps 3 and 4 of the RECONSTRUCT method and show that for a large $\mathbb{F}$, w.h.p., these steps determine the needed partition correctly.

Let Φ be a random multilinear formula sampled using $\texttt{SAMPLE}(X, \mathbb{F})$ and let $\hat{\Phi} = A(\bar{v}, \bar{u})B(\bar{x}, \bar{y}) + C(\bar{v}, \bar{x})D(\bar{u}, \bar{y})$. In Step 3 of the RECONSTRUCT method, one chooses an m-sized subset S of X randomly and, in $\hat{\Phi}$, instantiates the variables in $X \setminus S$ to random values over $\mathbb{F}$ to get $\hat{\Phi}_S = A_S(\bar{v}_S, \bar{u}_S)B_S(\bar{x}_S, \bar{y}_S) + C_S(\bar{v}_S, \bar{x}_S)D_S(\bar{u}_S, \bar{y}_S)$. Using Chernoff's bound it easily follows that w.h.p sizes of the sets $\bar{v}_S$, etc., are $\Omega(m)$. Let $Y = S$ and $Z = X \setminus S$. In the SAMPLE method, partitioning the set $Y \cup Z$ at a $\times$ gate (where $|Y| \leq |Z|$) into two equal-sized sets $\{\bar{a}\}, \{\bar{b}\}$ can be viewed as follows: label the y_i's in Y with independent uniform 0-1 values, including the y_i's with label 0 in $\{\bar{a}\}$ and label 1 in $\{\bar{b}\}$, and finally, place the Z variables randomly to make $|\bar{a}| = |\bar{b}|$. It is now easy to see that in the above expression of $\hat{\Phi}_S$, the polynomials A_S, B_S, C_S, D_S are close in distribution to a multilinear formula sampled using the following sampling method on their respective variable sets.

Sampling Method $\texttt{SAMPLE}_2(X, \mathbb{F})$**:**

Step 1: $\Psi \leftarrow \texttt{CONSTRUCT}_2(X, +)$.

Step 2: Let W be the set of wires in Ψ incident to a + gate. Let Φ be the syntactic multilinear arithmetic formula obtained by labeling each $w_i \in W$ by a randomly and independently chosen $c_i \in_R \mathbb{F}$.

Step 3: **return**(Φ)

where $\texttt{CONSTRUCT}_2(X, op)$:

Case 1: $X = \{x_i\}$. Let Ψ be the formula with a + gate at the root that has one wire incident to it from x_i and one from the field element 1.

Case 2: $op = \times$. Label each $x_i \in X$ with independent uniformly chosen 0-1 values. Include the x_i's labeled 0 in a set X_1 and the rest in X_2. If some X_i is empty then repeat. Let $\Psi_1 \leftarrow \texttt{CONSTRUCT}_2(X_1, +)$, $\Psi_2 \leftarrow \texttt{CONSTRUCT}_2(X_2, +)$. Let Ψ be the formula with a $\times$ gate at the root and Ψ_1, Ψ_2 as its two children.

Case 3: $op = +$. Let $\Psi_1 \leftarrow \texttt{CONSTRUCT}_2(X, \times)$, $\Psi_2 \leftarrow \texttt{CONSTRUCT}_2(X, \times)$. Let Ψ be the formula with a + gate at the root and Ψ_1, Ψ_2 as its two children.

Step: **return**(Ψ)

Theorem 5.6 (Uniqueness of Partition)**.** *Let $\{\{\bar{a}\}, \{\bar{b}\}\}$ and $\{\{\bar{c}\}, \{\bar{d}\}\}$ be partitions of $\{\bar{y}\} \cup \{\bar{z}\}$ and $|\bar{a}|$, $|\bar{b}|$, $|\bar{c}|$, $|\bar{d}|$, $|\bar{y}|$, $|\bar{z}|$ are all $\Omega(m)$. Let $A(\bar{a})$, $B(\bar{b})$, $C(\bar{c})$, $D(\bar{d})$ be polynomials independently computed by random multilinear formulas sampled using $\texttt{SAMPLE}_2$ over the indicated variable sets and field $\mathbb{F}$. Then for independent $\alpha, \beta \in_R \mathbb{F}$,*

$$\Pr[\text{Rank}_{\{\bar{y}\}\{\bar{z}\}}(\alpha \cdot AB + \beta \cdot CD) \leq 2] \leq \frac{2^{O(m)}}{|\mathbb{F}|} + \frac{1}{2^{\Omega(m)}},$$

unless

1. *either $\{\bar{y}\} = \{\bar{a}\}$ & $\{\bar{z}\} = \{\bar{b}\}$ or $\{\bar{y}\} = \{\bar{b}\}$ & $\{\bar{z}\} = \{\bar{a}\}$, and*
2. *either $\{\bar{y}\} = \{\bar{c}\}$ & $\{\bar{z}\} = \{\bar{d}\}$ or $\{\bar{y}\} = \{\bar{d}\}$ & $\{\bar{z}\} = \{\bar{c}\}$*

Before we sketch a proof of Theorem 5.6 (full proof appears in Appendix B), let's see how it is used in the proof of Theorem 2.2. In Step 3 of RECONSTRUCT, we consider the ranks of the partial derivative matrices for $\hat{\Phi}_S|_{\bar{v}',\bar{y}'}$ and $\hat{\Phi}_S|_{\bar{u}',\bar{x}'}$ w.r.t. partitions $\{\bar{v}',\bar{y}'\}$ and $\{\bar{u}',\bar{x}'\}$, respectively. First, note that if $\bar{v}'$ etc are the correct partition of S, i.e., in Φ_S, $v_S = \bar{v}'$ etc., then both the above matrices have rank at most 2. We use Theorem 5.6 to show that, w.h.p., the only partition of S (into four parts) that satisfy these two rank conditions is the correct partition. Indeed, by the discussion preceding Theorem 5.6, we can see that $A_S|_{\bar{v}',\bar{y}'}$, $B_S|_{\bar{v}',\bar{y}'}$, $C_S|_{\bar{v}',\bar{y}'}$, and $D_S|_{\bar{v}',\bar{y}'}$ can be viewed as samples from $\texttt{SAMPLE}_2$ on the variable set $\{\bar{v}',\bar{y}'\}$ (assigning $S \setminus \{\bar{v}' \cup \bar{y}'\}$ to random values). Similarly for $A_S|_{\bar{u}',\bar{x}'}$, etc., on $\{\bar{u}',\bar{x}'\}$. Now, Theorem 5.6 says if $\text{Rank}_{\{\bar{v}'\}\{\bar{y}'\}}(\hat{\Phi}_S|_{\bar{v}',\bar{y}'}) \leq 2$, then, w.h.p., the variables that $A_S|_{\bar{v}',\bar{y}'}$ etc. depend on must each be either $\bar{v}'$ and $\bar{y}'$. Thus, w.l.o.g., we must have $\bar{v}_S = \bar{v}'$ and $\bar{y}_S = \bar{y}'$. By a similar argument applied to $\hat{\Phi}_S|_{\bar{u}',\bar{x}'}$, we can conclude that $\bar{u}_S = \bar{u}'$ and $\bar{x}_S = \bar{x}'$. Note that since AB and CD are defined on two independent partitions of X, it is unlikely that A and C depend on the same set of variables. Applying this argument repeatedly for the seed partition augmented with x_i, we can also see that Step 4 associates each x_i with correct block of the seed partition. This concludes the proof that Steps 3 and 4 determine the correct partition for Φ.

Proof sketch for Theorem 5.6: Appendix B is dedicated to a detailed proof of this theorem. We first show, in Lemma B.1, that a random linear combination $\alpha f + \beta g$ has rank ≤ 2 w.r.t. a partition (Y, Z) of the underlying variable set only under very special conditions. The most natural of these is when f and g are both of rank 1, i.e., $f(Y,Z) = f_1(Y) \cdot f_2(Z)$ and $g(Y,Z) = g_1(Y) \cdot g_2(Z)$. The other (degenerate) conditions arise when at least one of f or g has rank 2 and can be categorized into a small number of special cases. The second part of the proof is to show that when $f = AB$ and $g = CD$ and A, B, C, and D are samples from $\texttt{SAMPLE}_2$, the degenerate conditions are satisfied with very low probability. This will imply AB and CD must satisfy the natural condition and hence their supports must satisfy (1) and (2). For the second part, we use two main arguments about a random formula according to $\texttt{SAMPLE}_2$ on m variables: (i) it must have rank at least two, w.h.p., for any nontrivial partition of its variables (Irreducibility Lemma, Lemma B.3) and (ii) for any partition (Y, Z) with $|Y|, |Z| \geq \Omega(m)$, it must contain many monomials in Z variables whose coefficients (which are polynomials in Y) must also contain many monomials in Y variables (Lemma B.4). By (i), we only need to consider when, say f, is of rank-2 w.r.t. some partition (not necessarily (Y, Z)). This, combined with any of the degeneracy conditions, implies that the number of statistically independent monomials in Y variables in the coefficient of a suitably chosen Z-monomial in g must be small (since they are determined by linear combinations given by the degeneracy conditions of a small number of coefficients of f's factors). But this contradicts (ii) since (by Lemma B.2) there must be many independent monomials in Y variables. ■

References

[Hås90] Johan Håstad. Tensor rank is np-complete. *J. Algorithms*, 11(4):644–654, 1990.

[Kal89] Erich Kaltofen. Factorization of polynomials given by straight-line programs. In *Randomness and Computation*, pages 375–412. JAI Press, 1989.

[KS01] Adam Klivans and Daniel A. Spielman. Randomness efficient identity testing of multivariate polynomials. In *STOC*, pages 216–223, 2001.

[KS06] Adam R. Klivans and Amir Shpilka. Learning restricted models of arithmetic circuits. *Theory of Computing*, 2(1):185–206, 2006.

[KS09] Zohar Shay Karnin and Amir Shpilka. Reconstruction of generalized depth-3 arithmetic circuits with bounded top fan-in. In *IEEE Conference on Computational Complexity*, pages 274–285, 2009.

[Nis91] Noam Nisan. Lower bounds for non-commutative computation. In *Proceedings of the twenty-third annual ACM symposium on Theory of computing*, STOC '91, pages 410–418, New York, NY, USA, 1991. ACM.

[Raz09] R. Raz. Multi-linear formulas for permanent and determinant are of super-polynomial size. *Journal of the Association for Computing Machinery*, 56(2), 2009.

[Shp09] Amir Shpilka. Interpolation of depth-3 arithmetic circuits with two multiplication gates. *SIAM J. Comput.*, 38(6):2130–2161, 2009.

[SY10] Amir Shpilka and Amir Yehudayoff. Arithmetic circuits: A survey of recent results and open questions. *Foundations and Trends in Theoretical Computer Science*, 5(3-4):207–388, 2010.

A Proofs for Section 5

A.1 Proof of Lemma 5.1

Lemma restated: Let $\Phi \sim \mathcal{D}(X, \mathbb{F})$. Then, for all nodes of Φ, the following hold with probability at least $1 - \frac{2^{O(n)}}{|\mathbb{F}|} - \frac{1}{n^{\Omega(1)}}$:

1. The polynomial computed by a node at (multiplicative) depth h is a homogenous polynomial of degree $\frac{n}{\beta_n 2^h}$.
2. The polynomial computed at a + gate is of the form $\alpha.A(\bar{v},\bar{u})B(\bar{x},\bar{y}) + \beta.C(\bar{v},\bar{x})D(\bar{u},\bar{y})$ where for all $\bar{p} \in \{\bar{v},\bar{u},\bar{x},\bar{y}\}$, $|\bar{p}| \geq \frac{1}{8}|\{\bar{v} \cup \bar{u} \cup \bar{x} \cup \bar{y}\}|$.
3. In the above polynomial computed at a + gate, for all $R \in \{A, B, C, D\}$, say $R(\bar{p},\bar{q})$, $R(\bar{0},\bar{q}) \neq 0$ and $R(\bar{p},\bar{0}) \neq 0$.

Proof. (1) The proof is by induction on depth. The polynomial computed at a + gate at depth h has the form $g(X_h) = \alpha.A(\bar{a})B(\bar{b}) + \beta.C(\bar{c})D(\bar{d})$ where A,B,C,D are sampled using `SAMPLE` on their respective variable sets and $|\bar{a}| = |\bar{b}| = |\bar{c}| = |\bar{d}| = |X_h|/2$ where $|X_h| = n/2^h$. Also, $\bar{a}, \bar{b}$ are disjoint and $\bar{c}, \bar{d}$ are disjoint. By induction, if A,B,C,D are homogenous polynomials of degree $d/2$ then, with probability $1 - 1/|\mathbb{F}|$, g would be a degree-d homogenous polynomial. We have the following expression where $\deg(m)$ denotes the degree of a node at depth h with a variable set of size m and by construction $\deg(\beta_n) = 1$.

$$\deg(m) = 2.(\deg(m/2)) \ldots = 2^t.\deg(m/2^t)$$

Hence, $\deg(m) = m/\beta_n$. As for a node at depth h we have $m = n/2^h$, part (1) follows. The probability bound follows from union bound.

(2) The polynomial computed at a + gate at depth h has the form $g(X_h) = A(\bar{a})B(\bar{b}) + C(\bar{c})D(\bar{d})$ where $\bar{a}, \bar{b}, \bar{c}, \bar{d}$ satisfy the properties stated in part (1). Now, let $\{\bar{v}\} = \{\bar{a}\} \cap \{\bar{c}\}$, $\{\bar{u}\} = \{\bar{a}\} \cap \{\bar{d}\}$, $\{\bar{x}\} = \{\bar{b}\} \cap \{\bar{c}\}$, $\{\bar{y}\} = \{\bar{b}\} \cap \{\bar{d}\}$. As the partition, $\{\{\bar{c}\},\{\bar{d}\}\}$ is chosen independent of $\{\{\bar{a}\},\{\bar{b}\}\}$ fix $\{\bar{a}\} = Y$ and $\{\bar{b}\} = Z$. Now choosing a random $\{\{\bar{c}\},\{\bar{d}\}\}$ can be viewed as labeling the y_i's in Y with independent uniform 0-1 values. Then including the y_i's, with label 0, in $\{\bar{c}\}$ else in $\{\bar{d}\}$. Then, placing the Z variables randomly to make the sizes of both sets equal. Hence, $|\bar{v}| = |\{\bar{c}\} \cap Y| = \zeta_1 + \zeta_2 + \ldots + \zeta_{|Y|}$ where ζ_i's are i.i.d 0-1 r.v.'s and $|\bar{u}| = |\{\bar{d}\} \cap Y| = |Y| - |\{\bar{c}\} \cap Y|$. Using Chernoff's bound, we have

$$\Pr[\ |\bar{v}| < |Y|/4 \text{ OR } |\bar{u}| < |Y|/4\ ] \ \leq\ 2^{-\delta|Y|},$$

for some constant $\delta > 0$ (e.g., $\delta = 1/128$). As, $|\bar{x}| = |\{\bar{c}\} \cap Z| = |\bar{c}| - |\{\bar{c}\} \cap Y| = |Y| - |\{\bar{c}\} \cap Y|$ it follows that,

$$\Pr[\ |\bar{v}| < |Y|/4 \text{ OR } |\bar{u}| < |Y|/4 \text{ OR } |\bar{x}| < |Y|/4 \text{ OR } |\bar{y}| < |Y|/4\ ] \ \leq\ 2 \cdot 2^{-\delta|Y|} \ \leq\ \frac{1}{2^{\beta_n)}} \leq 2^{-\Omega(\log^3 n)},$$

as in $\texttt{SAMPLE}(X, \mathbb{F})$, the variable set at any node of Φ is of size at least $\beta_n = \Theta(\log^3 n)$. Now, as there are $n^{O(1)}$ number of nodes, the stated probability bound follows from the union bound.

(3) From part (2), polynomial computed at a + gate is of the form $\alpha.A(\bar{v},\bar{u})B(\bar{x},\bar{y}) + \beta.C(\bar{v},\bar{x})D(\bar{u},\bar{y})$ where for all $\bar{p} \in \{\bar{v},\bar{u},\bar{x},\bar{y}\}$, $|\bar{p}| \geq \frac{1}{8}|\{\bar{v} \cup \bar{u} \cup \bar{x} \cup \bar{y}\}|$ and A,B,C,D are sampled using `SAMPLE` on their respective variable sets. For this part to follow it is enough to show that, in a polynomial g computed by a random formula sampled using `SAMPLE` on a n-sized variable set $Y \dot{\cup} Z$, with the stated probability, there is a monomial only on the Y variables. Also, w.l.o.g., $|Y| \leq |Z|$ and $|Y|$ is at least $n/8$. Proof will be by induction on the depth of g. Let $g = A'(\bar{a}')B'(\bar{b}') + C'(\bar{c}')D'(\bar{d}')$. Note that the number of monomials in only Y variables in $A'B'$ is product of the number of such monomials in A' and in B'. Moreover, the probability that in some step of the induction, these monomials will be canceled is at most $\frac{2^{O(n)}}{|\mathbb{F}|}$. Let $\delta := 1/\log n$. Now using Chernoff's bound,

$$\Pr[\ |\{\bar{a}'\} \cap Y| < |Y|(1-\delta)/2 \text{ OR } |\{\bar{b}'\} \cap Y| < |Y|(1-\delta)/2\ ] \leq 2^{-c \cdot \delta^2 \cdot |Y|}, \tag{3}$$

for some constant $c > 0$.

In the worst case, $|\{\bar{a}'\} \cap Y| = |Y|(1-\delta)/2$. Applying induction and assuming worst case every time we partition, we have the following bound for the number of monomials, denoted $M(|Y' \cup Z'|, |Y'|)$, in only Y' variables, in a polynomial over a set $Y' \dot{\cup} Z'$ with $|Y'| = \min\{|Y'|, |Z'|\}$, computed by a random formula:

$$M(n, |Y|) \geq (M(n/2, |Y|(1-\delta)/2)^2 \ldots \geq \left(M(n/2^h, |Y|(1-\delta)^h/2^h)\right)^{2^h}$$

For $2^h \leq n/\beta_n$, we have $|Y|(1-\delta)^h/2^h \geq 1$, and $M(n, |Y|) \geq M(\beta_n, 1)$. Since, by construction, for $|Y' \cup Z'| = \beta_n$ the formula will be a linear form with at least one term in $|Y'|$-variables, the lemma follows. Also we have ensured that at every step of induction $|Y'| \geq |Y|(1-\delta)^h/2^h = \Omega(\beta_n)$. Using this and $\delta = 1/\log n$ in inequality (3), the probability bound also follows. ■

A.2 Proof of Lemma 5.3

Lemma restated: Let f and g be two multilinear polynomials over the variable set $Y \cup Z$ and field $\mathbb{F}$. Then for any $S \subset \mathbb{F}$, and two independent random variables α, β,

$$\Pr_{\alpha,\beta \in_R S}[\text{Rank}_{YZ}(\alpha.f + \beta.g) \geq \max\{\text{Rank}_{YZ}(f), \text{Rank}_{YZ}(g)\}] \geq 1 - \frac{2^{\min\{|Y|,|Z|\}}}{|S|}.$$

Proof is an immediate consequence of the following lemma.

Lemma A.1. *Let M_1 and M_2 be two $r \times r$ matrices over a field $\mathbb{F}$, such that M_1 has a full rank. Then for any $S \subset \mathbb{F}$, and two independent random variables α, β,*

$$\Pr_{\alpha,\beta \in_R S}[\ \alpha.M_1 + \beta.M_2 \text{ has full rank }\ ] \geq 1 - \frac{r}{|S|}.$$

Proof. The matrix $\alpha.M_1 + \beta.M_2$ has a full rank iff it has a non-zero determinant. Using induction on r, one can easily see that $\det(\alpha.M_1 + \beta.M_2)$ is a degree r polynomial in α with coefficient of α^r equal to $\det(M_1)$ and hence non-zero as M_1 has full rank. For any choice of β, the said degree r polynomial in α can have at most r roots. Hence the probability that $\det(\alpha.M_1 + \beta.M_2) = 0$ is at most $r/|S|$. ■

A.3 Proof of Lemma 5.4

Lemma restated: Let $\mathbb{F}$ be a field with at least $d+1$ elements and let $f \in \mathbb{F}[x_1, \ldots, x_n]$ be a degree d polynomial. Given blackbox access to f we can simulate blackbox access to $f^{[r]}$'s, where $f^{[r]}$ denotes the homogenous degree-r part of f.

Proof. To determine $f^{[r]}(\bar{\beta})$ for a given $\bar{\beta} \in \mathbb{F}^n$ query $f(\bar{\beta}), f(2\bar{\beta}), \ldots, f((d+1)\bar{\beta})$ to the oracle, where for $\bar{\beta} = (\beta_1, \ldots, \beta_n)$, $i\bar{\beta}$ denotes $(i\beta_1, \ldots, i\beta_n)$. Then we have,

$$f(i\bar{\beta}) = i^0 f^{[0]}(\bar{\beta}) \ + \ i f^{[1]}(\bar{\beta}) \ + \ \ldots \ + \ i^d f^{[d]}(\bar{\beta}),$$

or,

$$\begin{bmatrix} f(\bar{\beta}) \\ f(2\bar{\beta}) \\ \vdots \\ f((d+1)\bar{\beta}) \end{bmatrix} = \begin{bmatrix} 1 & 1 & \ldots & 1^d \\ 1 & 2 & \ldots & 2^d \\ \vdots & \vdots & \ddots & \vdots \\ 1 & d+1 & \ldots & (d+1)^d \end{bmatrix} \begin{bmatrix} f^{[0]}(\bar{\beta}) \\ f^{[1]}(\bar{\beta}) \\ \vdots \\ f^{[d]}(\bar{\beta}) \end{bmatrix}.$$

As the above coefficient matrix of $f^{[r]}(\bar{\beta})$'s is a vandermonde matrix(and hence invertible), $f^{[r]}(\bar{\beta})$ can be easily determined. ■

A.4 Algorithm TRICKLEDOWN

Theorem 5.5 restated: Let $\{\{\bar{v}\},\{\bar{u}\},\{\bar{x}\},\{\bar{y}\}\}$ be a partition of $\{x_1,\ldots,x_n\}$ and $f(\bar{v},\bar{u},\bar{x},\bar{y}) = A(\bar{v},\bar{u})B(\bar{x},\bar{y}) + C(\bar{v},\bar{x})D(\bar{u},\bar{y})$ be a non-zero polynomial such that,

1. A,B,C,D are homogenous multilinear polynomials over the indicated variable sets,
2. either $\deg(AB) \neq \deg(CD)$ or $\deg(A) = \deg(B) = \deg(C) = \deg(D)$,
3. for all $R \in \{A,B,C,D\}$, say $R(\bar{p},\bar{q})$, $R(\bar{0},\bar{q}) \neq 0$ and $R(\bar{p},\bar{0}) \neq 0$.
4. for all $\bar{p} \in \{\bar{v},\bar{u},\bar{x},\bar{y}\}$, $|\bar{p}| \geq \delta n$, for some $\delta > 0$

Proof. The proof follows from the algorithm TRICKLEDOWN below.
Input: The partition $\{\{\bar{v}\},\{\bar{u}\},\{\bar{x}\},\{\bar{y}\}\}$ and an oracle for $A(\bar{v},\bar{u})B(\bar{x},\bar{y})+C(\bar{v},\bar{x})D(\bar{u},\bar{y})$ where A,B,C,D are polynomials satisfying the above stated properties.
Output: Blackboxes for A,B,C,D.
Algorithm: TRICKLEDOWN

Step 1: Using blackbox for $f = AB + CD$, construct blackboxes for $f^{[i]}$'s for all $i \in [n]$.

Step 2: For $i \in [n]$, using blackbox PIT, determine if $f^{[i]} \neq 0$. If there is only one such i then proceed to the next step. Otherwise let $f^{[i]}$, $f^{[j]}$ be non-zero. For $f^{[i]}(\bar{v},\bar{u},\bar{x},\bar{y})$, determine using blackbox PIT, if $f^{[i]}(\bar{v},\bar{u},\bar{0},\bar{0})$ is 0. If yes, conclude $f^{[i]} = AB$ and $f^{[j]} = CD$, else the other way. Using Kaltofen's factoring algorithm, construct blackboxes for irreducible factors of $A(\bar{v},\bar{u})B(\bar{x},\bar{y})$. For each factor $h(\bar{v},\bar{u},\bar{x},\bar{y})$, determine, using blackbox PIT, if $h(\bar{0},\bar{0},\bar{x},\bar{y})$ is 0. If yes, conclude it is a factor of A else B. Similarly, construct blackboxes for C and D.

Step 3: Determining degrees of A,B,C,D. Using Kaltofen's factoring algorithm, gain blackbox access to irreducible factors of $f(\bar{v},\bar{u},\bar{0},\bar{0}) = C(\bar{v},\bar{0})D(\bar{u},\bar{0})$. For each factor h, determine, using blackbox PIT, if h becomes the zero polynomial after instantiating $\bar{v}$ to $\bar{0}$. If yes it is a factor of $C(\bar{v},\bar{0})$ else $D(\bar{u},\bar{0})$. Similarly, construct blackboxes for $C(\bar{0},\bar{x})$ and $D(\bar{0},\bar{y})$. Having constructed blackboxes for $C(\bar{v},\bar{0})$ and $D(\bar{u},\bar{0})$, conclude $d = \deg(C) = \log\left(\frac{C(2\bar{\alpha},\bar{0})}{C(\bar{\alpha},\bar{0})}\right)$ for a randomly chosen $\bar{\alpha} \in \mathbb{F}^{|\bar{v}|}$, and similarly for D, A, B.

Step 4: Constructing blackbox for C. To determine $C(\bar{\alpha},\bar{\beta})$, for any $\bar{\alpha} \in \mathbb{F}^{|\bar{v}|}$, $\bar{\beta} \in \mathbb{F}^{|\bar{x}|}$, substitute $\bar{x} = \bar{\beta}$ and $\bar{y} = \bar{\gamma}$ for $\bar{\gamma} \in_R \mathbb{F}^{|\bar{y}|}$. Then,

$$f(\bar{v},\bar{u},\bar{\beta},\bar{\gamma}) = \underbrace{A(\bar{v},\bar{u})B(\bar{\beta},\bar{\gamma})}_{\text{only degree } \deg(A) \text{ terms}} + \underbrace{C(\bar{v},\bar{\beta})D(\bar{u},\bar{\gamma})}_{\text{terms can have degree} > \deg(A)} = A(\bar{v},\bar{u})B(\bar{\beta},\bar{\gamma}) + \hat{C}(\bar{v})\hat{D}(\bar{u}).$$

Let $g^{[d]}$ denote the homogenous degree-d part of g. Then,

$$\hat{C}(\bar{v}) = \hat{C}^{[d]}(\bar{v}) \ldots + \hat{C}^{[1]}(\bar{v}) + C(\bar{0},\bar{\beta}) \text{ and } \hat{D}(\bar{u}) = \hat{D}^{[d]}(\bar{u}) \ldots + \hat{D}^{[1]}(\bar{u}) + D(\bar{0},\bar{\gamma}).$$

Note that $f(\bar{v},\bar{u},\bar{\beta},\bar{\gamma})^{[2d]} = C^{[d]}(\bar{v})D^{[d]}(\bar{u})$. Using Kaltofen's algorithm, obtain blackboxes for $\hat{C}^{[d]}(\bar{v})$ and $\hat{D}^{[d]}(\bar{u})$ using blackbox for $f(\bar{v},\bar{u},\bar{\beta},\bar{\gamma})^{[2d]}$. As Kaltofen's algorithm gives blackboxes for irreducible factors of $C^{[d]}(\bar{v})D^{[d]}(\bar{u})$ and any such factor depends on either $\bar{v}$ or $\bar{u}$, to find out if $h(\bar{v},\bar{u})$ depends on $\bar{u}$ use blackbox PIT on $h(\bar{v},\bar{0})$.

Step 5: Constructing blackboxes for $\hat{C}^{[i]}(\bar{v})$ and $\hat{D}^{[i]}(\bar{u})$ for $i \in [d-1]$. Having gained blackboxes for $\hat{C}^{[d]}(\bar{v})$ and $\hat{D}^{[d]}(\bar{u})$ we note that,

$$f(\bar{v},\bar{u},\bar{\beta},\bar{\gamma})^{[2d-1]} = \hat{C}^{[d]}(\bar{v})\hat{D}^{[d-1]}(\bar{u}) + \hat{C}^{[d-1]}(\bar{v})\hat{D}^{[d]}(\bar{u}) \quad (4)$$

$$\implies f(\bar{v},2\bar{u},\bar{\beta},\bar{\gamma})^{[2d-1]} = 2^{d-1}\hat{C}^{[d]}(\bar{v})\hat{D}^{[d-1]}(\bar{u}) + 2^d\hat{C}^{[d-1]}(\bar{v})\hat{D}^{[d]}(\bar{u}) \quad (5)$$

$$\implies \frac{1}{2^{d-1}}f(\bar{v},2\bar{u},\bar{\beta},\bar{\gamma})^{[2d-1]} = \hat{C}^{[d]}(\bar{v})\hat{D}^{[d-1]}(\bar{u}) + 2\hat{C}^{[d-1]}(\bar{v})\hat{D}^{[d]}(\bar{u}) \quad (6)$$

From (3) − (1) we have,

$$\hat{C}^{[d-1]}(\bar{v}) \;=\; \frac{1}{\hat{D}^{[d]}(\bar{u})}\left[\frac{1}{2^{d-1}}f(\bar{v},2\bar{u},\bar{\beta},\bar{\gamma})^{[2d-1]} - f(\bar{v},\bar{u},\bar{\beta},\bar{\gamma})^{[2d-1]}\right]$$

As we have blackbox access to $f^{[2d-1]}$ and $\hat{D}^{[d]}(\bar{u})$, we have blackbox access to $\hat{C}^{[d-1]}(\bar{v})$ after instantiating $\bar{u}$ randomly to avoid making the denominator vanish in the above equation. Similarly we have blackbox access to $\hat{D}^{[d-1]}(\bar{u})$. In general, after constructing blackboxes for $\hat{C}^{[r]}(\bar{v})$,$\hat{D}^{[r]}(\bar{u})$ for all $r \in [d'+1:d]$, blackbox for $\hat{C}^{[d']}(\bar{v})$ can be constructed as follows

$$f(\bar{v},\bar{u},\bar{\beta},\bar{\gamma})^{[d+d']} = \hat{C}^{[d']}(\bar{v})\hat{D}^{[d]}(\bar{u}) + \left(\sum_{i=d'+1}^{d-1}\hat{C}^{[i]}(\bar{v})\hat{D}^{[d+d'-i]}(\bar{u})\right) + \hat{C}^{[d]}(\bar{v})\hat{D}^{[d']}(\bar{u})$$

$$\frac{f(\bar{v},2\bar{u},\bar{\beta},\bar{\gamma})^{[d+d']}}{2^{d'}} = 2^{d-d'}\hat{C}^{[d']}(\bar{v})\hat{D}^{[d]}(\bar{u}) + \left(\sum_{i=d'+1}^{d-1}2^{d-i}\hat{C}^{[i]}(\bar{v})\hat{D}^{[d+d'-i]}(\bar{u})\right) + \hat{C}^{[d]}(\bar{v})\hat{D}^{[d']}(\bar{u})$$

Subtracting the two equations we have,

$$\hat{C}^{[d']}(\bar{v}) = \frac{2^{d'}}{(2^d-2^{d'})\hat{D}^{[d]}(\bar{u})}\left[\frac{f(\bar{v},2\bar{u},\bar{\beta},\bar{\gamma})^{[d+d']}}{2^{d'}} - f(\bar{v},\bar{u},\bar{\beta},\bar{\gamma})^{[d+d']} - \sum_{i=d'+1}^{d-1}(2^{d-i}-1)\hat{C}^{[i]}(\bar{v})\hat{D}^{[d+d'-i]}(\bar{u})\right]$$

Hence, using the above procedure blackboxes for $\hat{C}^{[d']}(\bar{v})$, for all $d' \in [d]$, can be constructed. Also, using the blackbox for $C(\bar{0},\bar{x})$ constructed in Step 3 determine $C(\bar{0},\bar{\beta})$. This completes our blackbox for $C(\bar{v},\bar{\beta})$.

Step 6: Repeat the above 3 steps similarly with the correct parameters to construct blackboxes for A,B and D.

■

B Uniqueness of the Seed Partition

In this section, we discuss Steps 3 and 4 of the RECONSTRUCT method and show that for a large $\mathbb{F}$, w.h.p., these steps determine the needed partition correctly. Before we discuss these steps we first present some technical lemmas which would be helpful to us in estimating the success probability of the said steps. Proofs of these lemmas appear after the proof of Theorem B.5. Throughout this paper LI stands for "Linearly Independent" and LD for "Linearly Dependent."

Lemma B.1. *Let f and g be two multilinear polynomials over a n-sized variable set $Y \cup Z$ and field $\mathbb{F}$. Then for any $S \subset \mathbb{F}$, and two independent random variables α, β*

$$\Pr_{\alpha,\beta \in_R S}[\text{Rank}_{YZ}(\alpha.f + \beta.g) > 2] \geq 1 - \frac{2^n}{|S|}$$

unless f and g have one the following forms,

1. *$f = f_1(Y)f_2(Z)$ and $g = g_1(Y)g_2(Z)$*
2. *$f = f_1(Y)f_2(Z) + f_3(Y)f_4(Z)$ (f_1, f_3 are LI, f_2, f_4 are LI) and either $g = [a.f_1(Y) + b.f_3(Y)]g_2(Z)$ or $g = g_1(Y)[a.f_2(Z) + b.f_4(Z)]$*
3. *$f = f_1(Y)f_2(Z) + f_3(Y)f_4(Z)$ (f_1, f_3 are LI, f_2, f_4 are LI) and $g = [a.f_1(Y) + b.f_3(Y)]g_2(Z) + [c.f_1(Y) + d.f_3(Y)]g_4(Z)$ (g_2, g_4 are LI and $ad \neq bc$)*
4. *$f = f_1(Y)f_2(Z) + f_3(Y)f_4(Z)$ and $g = [a.f_1(Y) + b.f_3(Y)]g_2(Z) + g_3(Y)[c.f_2(Z) + d.f_4(Z)]$ (f_1, f_3, g_3 are LI, f_2, f_4, g_4 are LI and $ac = -bd$)*

and their analogous cases, where f_i's and g_i's are any multilinear polynomials on their indicated variable sets and $a, b, c, d \in \mathbb{F}$.

Lemma B.2. *Let S be a set of multilinear monomials over $\{r_1, r_2, \ldots, r_n\}$ where r_i's are independent r.v.'s and each $r_i \in_R \mathbb{F}^*$. Then for every $M \in S$ there exists a set $S_M \subset S$ such that*

1. *$|S_M| \geq \log |S| - 1$ and*
2. *$S_M \cup \{M\}$ is a set of independent uniform r.v.'s over $\mathbb{F}^*$.*

Placement of random field elements on the wires of a random multilinear formula in SAMPLE$(X, \mathbb{F})$ method where $X = \{x_1, x_2, \ldots, x_n\}$:

While sampling a multilinear formula from the set $\mathcal{M}(X, \mathbb{F})$ we first sampled a formula without any field elements using the method CONSTRUCT and later placed field elements, chosen independently and uniformly from $\mathbb{F}$, on its wires. Also note that, distinct wires originating from any of the x_i's, have distinct independent uniform r.v.'s on them. For instance consider a multilinear formula on X and that every x_i has at most one wire originating from it. Let the formula be $\sum_{k=1}^{N} \alpha_k.M_k$ where M_i's are multilinear monomials. Now for all x_i's, if we place a r.v. r_i on the wire from x_i then a term like $\alpha.x_1x_3x_n$ becomes $\alpha.r_1r_3r_n.x_1x_3x_n$. Hence essentially, the coefficient of a multilinear monomial M on X, is of the form $\alpha_M.M_r$ where M_r is the multilinear monomial $\Pi_{x_i \in M} r_i$ and each α_M is independent of r_i's. By Lemma B.2, for every monomial M there is a set of $\log N$ monomials containing M such that the set of coefficients of these monomials is mutually independent. Also, it is easy to note that this is true, even after instantiating variables to random values over $\mathbb{F}$.

Instantiating $n-m$ variables to random field elements in Step 3 of the RECONSTRUCT method:

Let Φ be a random multilinear formula sampled using $\texttt{SAMPLE}(X,\mathbb{F})$ and let $\hat{\Phi} = A(\bar{v},\bar{u})B(\bar{x},\bar{y}) + C(\bar{v},\bar{x})D(\bar{u},\bar{y})$. In Step 3 of the RECONSTRUCT method, one chooses a m-sized subset S of X randomly and ,in $\hat{\Phi}$, instantiates the variables in $X\backslash S$ to random values over $\mathbb{F}$ to get $\hat{\Phi}_S = A_S(\bar{v}_S,\bar{u}_S)B_S(\bar{x}_S,\bar{y}_S) + C_S(\bar{v}_S,\bar{x}_S)D_S(\bar{u}_S,\bar{y}_S)$. Using Chernoff's bound it easily follows that w.h.p sizes of the sets $\bar{r}_S$'s are $\Omega(m)$. Let $Y = S$ and $Z = X\backslash S$. In the SAMPLE method, partitioning a set $Y \cup Z$ on a $\times$ gate(where $|Y| \leq |Z|$) into two equal sized sets $\{\bar{a}\},\{\bar{b}\}$ can be viewed as labelling the y_i's in Y with independent uniform 0-1 values. Then including the y_i's, with label 0, in $\{\bar{a}\}$ else in $\{\bar{b}\}$. Then, placing the Z variables randomly to make the sizes of both sets equal. Hence in the above expression of $\hat{\Phi}_S$, the polynomials A_S, B_S, C_S, D_S are close in distribution to a multilinear formula sampled using the following sampling method on their respective varibale sets.

Sampling Method : $\texttt{SAMPLE}_2(X,\mathbb{F})$:

Step 1: $\Psi \leftarrow \texttt{CONSTRUCT}_2(X,+)$

Step 2: Let W be the set of wires in Ψ incident to a + gate. For each $w_i \in W$ place c_i on w_i to get a formula Φ, where each $c_i \in_R \mathbb{F}$ and c_i's are sampled independently.

Step 3: **return**(Φ)

where $\texttt{CONSTRUCT}_2(X,op)$:

Case 1: $X = \{x_i\}$. Let Ψ be the formula with a + gate at the root that has one wire incident to it from x_i and one from the field element 1.

Case 2: $op = \times$. Label each $x_i \in X$ with independent uniformly chosen 0-1 values. Include the x_i's labeled 0 in a set X_1 and the rest in X_2. If some X_i is empty then repeat. Let $\Psi_1 \leftarrow \texttt{CONSTRUCT}_2(X_1,+)$, $\Psi_2 \leftarrow \texttt{CONSTRUCT}_2(X_2,+)$. Let Ψ be the formula with a $\times$ gate at the root and Ψ_1,Ψ_2 as its two children.

Case 3: $op = +$. Let $\Psi_1 \leftarrow \texttt{CONSTRUCT}_2(X,\times)$, $\Psi_2 \leftarrow \texttt{CONSTRUCT}_2(X,\times)$. Let Ψ be the formula with a + gate at the root and Ψ_1,Ψ_2 as its two children.

Step: **return**(Ψ)

Lemma B.3 (Irreducibility Lemma)**.** *Let f_R be the polynomial computed by a random multilinear formula over the variables set $X = \{x_1, x_2, \ldots, x_m\}$ and field $\mathbb{F}$ sampled using $\texttt{SAMPLE}_2$. The probability that there exists a proper partition $\{Y,Z\}$ of X such that $\text{Rank}_{YZ}(f_R) = 1$ is at most $\frac{2^{O(m)}}{|\mathbb{F}|}$.*

Lemma B.4. *Let $\{Y,Z\}$, with $|Y| \leq |Z|$, be a partition of variable set $X = \{x_1,\ldots,x_m\}$ such that both $|Y|$, $|Z|$ are at least γm for some $\gamma > 0$ and δ be a sufficiently large integer constant . Let f be the polynomial computed by a random multilinear formula sampled using $\texttt{SAMPLE}_2(X,\mathbb{F})$. Then, with probability $1 - \frac{2^{O(m)}}{|\mathbb{F}|} - \frac{1}{2^{\gamma m/18\log^2\delta}}$*

1. *there are at least δ distinct monomials multilinear in Z variables such that coefficients of these are polynomials in Y each containing at least δ monomials and*
2. *$Rank_{YZ}(f) > 2$.*

Theorem B.5 (Uniqueness of Partition)**.** *Let $\{\{\bar{a}\},\{\bar{b}\}\}$ and $\{\{\bar{c}\},\{\bar{d}\}\}$ be partitions of $\{\bar{y}\}\cup\{\bar{z}\}$ and $|\bar{a}|$, $|\bar{b}|$, $|\bar{c}|$, $|\bar{d}|$, $|\bar{y}|$, $|\bar{z}|$ are all $\Omega(m)$. Let $A(\bar{a})$, $B(\bar{b})$, $C(\bar{c})$, $D(\bar{d})$ be polynomials independently computed by*

random multilinear formulas sampled using $\texttt{SAMPLE}_2$ *over the indicated variable sets and field* $\mathbb{F}$*. Then for independent* $\alpha, \beta \in_R \mathbb{F}$*,*

$$\Pr[\text{Rank}_{\{\bar{y}\}\{\bar{z}\}}(\alpha.AB + \beta.CD) \leq 2] \leq \frac{2^{O(m)}}{|\mathbb{F}|} + \frac{1}{2^{\Omega(m)}}$$

unless,

1. *either* $\{\bar{y}\} = \{\bar{a}\}$ & $\{\bar{z}\} = \{\bar{b}\}$ *or* $\{\bar{y}\} = \{\bar{b}\}$ & $\{\bar{z}\} = \{\bar{a}\}$*, and*
2. *either* $\{\bar{y}\} = \{\bar{c}\}$ & $\{\bar{z}\} = \{\bar{d}\}$ *or* $\{\bar{y}\} = \{\bar{d}\}$ & $\{\bar{z}\} = \{\bar{c}\}$

Proof. As the partition $\{\{\bar{y}\}, \{\bar{z}\}\}$ is clear in this context we would denote $\text{Rank}_{\{\bar{y}\}\{\bar{z}\}}$ by Rank. From Lemma 5.3, if any of A, B, C, D has Rank greater than 2, then indeed the above probability bound holds. Now if for some $\bar{r} \in \{\bar{a}, \bar{b}, \bar{c}, \bar{d}\}$ we have both $|\{\bar{r}\} \cap \{\bar{y}\}| = \Omega(m)$ and $|\{\bar{r}\} \cap \{\bar{z}\}| = \Omega(m)$ then by Lemma B.4, with the above probability, its respective random formula will have Rank greater than 2. W.l.o.g let $|\{\bar{a}\} \cap \{\bar{y}\}| = \Omega(m)$ and hence $|\{\bar{b}\} \cap \{\bar{z}\}| = \Omega(m)$. If either (1) or (2) doesn't hold then at least one of $\bar{a}$, $\bar{b}$, $\bar{c}$, $\bar{d}$ is such that it has at least one variable from both $\bar{y}$ and $\bar{z}$. W.l.o.g let $z_1 \in \{\bar{a}\}$. Using the Irreducibility lemma, we have $\text{Rank}(A) \geq 2$ with the above probability. Now if some $y_i \in \{\bar{b}\}$ then again the Irreducibility lemma would imply $\text{Rank}(B) \geq 2$ and hence $\text{Rank}(AB) \geq 4$. Hence let $\{\bar{b}\} \cap \{\bar{y}\} = \phi$. In the worst case $\{\bar{a}\} \cap \{\bar{z}\} = \{z_1\}$ (this will be easy to see from the following arguments) and hence w.l.o.g we have $\{\bar{a}\} = \{\bar{y}, z_1\}$ and $\{\bar{b}\} = \{z_2, \ldots\}$. Let $A(\bar{y}, z_1) = a_1(\bar{y})a_2(z_1) + a_3(\bar{y})a_4(z_1)$ with a_1, a_3 Linearly Independent (LI) and a_2, a_4 LI be a fixed representation of A. From Lemma B.1, the only way left for $\alpha.AB + \beta.CD$ to have Rank at most 2 are the following cases and we show that for any fixed A and B the following will not hold with the stated probability,

Case 2: $C(\bar{c})D(\bar{d}) = [p.a_1(\bar{y}) + q.a_3(\bar{y})]U(\bar{z})$ or $C(\bar{c})D(\bar{d}) = U(\bar{y})[p.a_2(z_1) + q.a_4(z_1)]B(z_2, \ldots)$ for some multilinear polynomial U possibly dependent on C and D and $p, q \in \mathbb{F}$.
If any of $\bar{c}$ or $\bar{d}$ has a variable of both $\bar{y}$ and $\bar{z}$ then by the Irreducibility lemma its Rank will be at least 2 and hence CD couldn't be equal to the RHS. Hence w.l.o.g $\{\bar{c}\} = \{\bar{y}\}$, $\{\bar{d}\} = \{\bar{z}\}$ and therefore $C(\bar{y}) = p.a_1(\bar{y}) + q.a_3(\bar{y})$. Here although p and q are possibly dependent on C, a_1 and a_3 are dependent only on A and hence are fixed. As a_1, a_3 are LI there exist two monomials M_1 and M_2 such that specifying the coefficients of M_1 and M_2 in $p.a_1(\bar{y}) + q.a_3(\bar{y})$ completely determine p and q and hence all its coefficients. But from Lemma B.4 we have that with the stated probability, there are 16 monomials in C and hence for any two monomials M_1 and M_2 in C, from Lemma B.2, there is a third one such the coefficient of this monomial is independent of that of the other two and hence w.h.p fixing the coefficients of these two monomials do not determine the LHS completely.
Similarly, the other case results in $D(\bar{z}) = [p.a_2(z_1) + q.a_4(z_1)]B(z_2, \ldots)$ with a_2, a_4, B fixed and again the same argument follows.

Case 3: $C(\bar{c})D(\bar{d}) = [p.a_1(\bar{y}) + q.a_3(\bar{y})]U_1(\bar{z}) + [r.a_1(\bar{y}) + s.a_3(\bar{y})]U_2(\bar{z})$ or
$C(\bar{c})D(\bar{d}) = U_1(\bar{y})[p.a_2(\bar{z_1}) + q.a_4(z_1)]B(z_2, \ldots) + U_2(\bar{y})[r.a_2(\bar{z_1}) + s.a_4(z_1)]B(z_2, \ldots)$ for some multilinear polynomials U_1 and U_2 possibly dependent on C and D. For the first subcase, from Lemma B.4 we have that w.h.p., on LHS there is a monomial M_z in $\{\bar{z}\}$ variables such the coefficient of M_z is a polynomial $g(\bar{y})$ having at least 16 monomials. Now comparing the coefficients of M_z on both sides we have, $g(\bar{y}) = (p.a_1(\bar{y}) + q.a_3(\bar{y}))u_1 + (r.a_1(\bar{y}) + s.a_3(\bar{y}))u_2$ where $p, q, r, s, u_1, u_2 \in \mathbb{F}$ are possibly dependent on LHS but a_1 and a_3 are fixed. This further implies that $g(\bar{y}) = (p.u_1 + r.u_2)a_1(\bar{y}) + (q.u_1 + s.u_2)a_3(\bar{y})$ and again the same argument as in the previous case follows.
Similarly for the other subcase we compare the coefficients of the monomial in $\{\bar{y}\}$ such that its coefficient is the polynomial having maximum number of monomials in $\{\bar{z}\}$.

Case 4: $C(\bar{c})D(\bar{d}) = [p.a_1(\bar{y}) + q.a_3(\bar{y})]U_1(\bar{z}) + U_2(\bar{y})[r.a_2(\bar{z_1}) - \frac{pr}{q}.a_4(z_1)]B(z_2, \ldots)$ for some multilinear polynomials U_1 and U_2 possibly dependent on C and D.
For the first subcase, from Lemma B.4 we have that w.h.p on LHS there is a monomial M_1 in $\{\bar{z}\}$ variables

such the coefficient of M_1 is a polynomial $\Delta_1 g_1(\bar{y})$ having δ monomials and where Δ_1 is a uniform r.v. over $\mathbb{F}$ that depends only on the r.v.'s placed on the wires of variables in M_1. Similarly let M_2 be any another monomial in $\{\bar{z}\}$ variables with coefficient $\Delta_2 g_2(\bar{y})$ where Δ_2 is independent of Δ_1. Now comparing the coefficients of M_1 and M_2 on both sides we have,

$$\Delta_1 g_1(\bar{y}) = [p.a_1(\bar{y}) + q.a_3(\bar{y})]u_1 + U_2(\bar{y})\psi_1(p,q,r) \ , \ \Delta_2 g_2(\bar{y}) = [p.a_1(\bar{y}) + q.a_3(\bar{y})]u_2 + U_2(\bar{y})\psi_2(p,q,r)$$

where $p, q, r, u_1, u_2 \in \mathbb{F}$ are possibly dependent on LHS, a_1 and a_3 are fixed and ψ_1 and ψ_2 are fixed functions of p, q, r. Eliminating $U_2(\bar{y})$ we have,

$$\Delta_1 g_1(\bar{y})\psi_2(p,q,r) - \Delta_2 g_2(\bar{y})\psi_1(p,q,r) = \psi_2(p,q,r)[p.a_1(\bar{y}) + q.a_3(\bar{y})]u_1 - \psi_1(p,q,r)[p.a_1(\bar{y}) + q.a_3(\bar{y})]u_2$$

But as p, q, r can depend on Δ_1 and Δ_2 the LHS may not have a large number of monomials and hence we cannot apply the previous argument directly. Here we note that we can compare the coeffcients of many monomials in $\{\bar{z}\}$ as from Lemma B.4 w.h.p there will be a at least δ such monomials. Hence we compare the coefficients of the monomials in $\{\bar{z}\}$ on LHS such that there corresponding Δ's are mutually independent and are independent to Δ_1. From Lemma B.2, there exists a set of $\log \delta$ such monomials and hence we get $\log \delta$ equations of the form,

$$\Delta_i g_i(\bar{y}) = [p.a_1(\bar{y}) + q.a_3(\bar{y})]u_i + U_2(\bar{y})\psi_i(p,q,r)$$

where ψ_i's are fixed functions and Δ_i's are mutually independent. Eliminating $U_2(\bar{y})$ from pairs of equations $((1),(i))$ we get $\log \delta - 1$ equations of the form,

$$\Delta_1 g_1(\bar{y})\psi_i(p,q,r) - \Delta_i g_i(\bar{y})\psi_1(p,q,r) = \psi_i(p,q,r)[p.a_1(\bar{y}) + q.a_3(\bar{y})]u_1 - \psi_1(p,q,r)[p.a_1(\bar{y}) + q.a_3(\bar{y})]u_i$$

As p, q, r can produce at most 8 uniform pairwise independent r.v.'s over $\mathbb{F}$, among $\log \delta$ Δ_i's (for a large enough δ) there is a Δ_j which is mutually independent to p, q, r and hence to all ψ_i's. Hence w.h.p the LHS of the following equation will have at least δ monomials and hence our previous argument will follow

$$\Delta_1.g_1(\bar{y})\psi_j(p,q,r) - \Delta_j.g_j(\bar{y})\psi_1(p,q,r) = \psi_j(p,q,r)[p.a_1(\bar{y}) + q.a_3(\bar{y})]u_1 - \psi_1(p,q,r)[p.a_1(\bar{y}) + q.a_3(\bar{y})]u_j$$

■

B.1 Proof of Lemma B.1

Proof. As the partition $\{Y, Z\}$ is clear in this context we would denote Rank_{YZ} by Rank. We now show that if none of the listed cases(and their analogs) hold then the stated probability bound holds. From Lemma 5.3 if any of f or g has Rank greater than 2 then indeed with above probability $\text{Rank}(\alpha.f+\beta.g) > 2$. Hence Case 1 occurs when both f, g have Rank 1. In rest of the cases at least one of f, g has Rank 2, w.l.o.g we assume f has Rank 2. Case 2 occurs when $\text{Rank}(g)$ is 1. Note that if Case 2 doesn't hold then, for $g = g_1(Y)g_2(Z)$, both f_1, f_3, g_1 would be LI and f_2, f_4, g_2 would be LI and hence Rank of their sum would be 3 with probability at least $1 - 2/|S|$. For Case 2 clearly the following sum has Rank at most 2,

$$\begin{aligned}\alpha.f + \beta.g &= \alpha[f_1(Y)f_2(Z) + f_3(Y)f_4(Z)] + \beta[a.f_1(Y) + b.f_3(Y)]g_2(Z) \\ &= f_1(Y)[\alpha.f_2(Z) + \beta.a.g_2(Z)] + f_3(Y)[\alpha.f_4(Z) + \beta.b.g_2(Z)]\end{aligned}$$

Cases 3 and 4 arise when $\text{Rank}(g)$ is also 2. Again from the previous argument, for $g = g_1(Y)g_2(Z) + g_3(Y)g_4(Z)$, if both f_1, f_3, g_1 are LI and f_2, f_4, g_2 are LI then Rank of their sum would be 3 with probability at least $1 - 2/|S|$ and similarly for g_3, g_4. Hence, for Rank of the sum to be 2, in both the summands in g, at least one of the factors must be Linearly Dependent (LD) on its counterparts in f. Therefore, this results in Case 3 and 4. In Case 3 the following sum has Rank at most 2,

$$\begin{aligned}\alpha.f + \beta.g &= \alpha[f_1(Y)f_2(Z) + f_3(Y)f_4(Z)] + \beta[\{a.f_1(Y) + b.f_3(Y)\}g_2(Z) + \{c.f_1(Y) + d.f_3(Y)\}g_4(Z)] \\ &= f_1(Y)[\alpha.f_2(Z) + \beta.a.g_2(Z) + \beta.c.g_4(Z)] + f_3(Y)[\alpha.f_4(Z) + \beta.b.g_2(Z) + \beta.d.g_4(Z)]\end{aligned}$$

In Case 4 the following sum has Rank at most 2 if $ac = -bd$,

$$\begin{aligned}\alpha f + \beta g &= \alpha[f_1(Y)f_2(Z) + f_3(Y)f_4(Z)] + \beta[\{a.f_1(Y) + b.f_3(Y)\}g_2(Z) + g_3(Y)\{c.f_2(Z) + d.f_4(Z)\}] \\ &= f_1(Y)[\alpha f_2(Z) + \beta.a.g_2(Z)] + f_3(Y)[\alpha f_4(Z) + \beta.b.g_2(Z)] + \beta g_3(Y)[c.f_2(Z) + d.f_4(Z)]\end{aligned}$$

The condition $ac = -bd$ arises as the coefficients $\alpha.f_2(Z) + \beta.a.g_2(Z)$, $\alpha.f_4(Z) + \beta.b.g_2(Z)$ and $\beta.c.f_2(Z) + \beta.d.f_4(Z)$ of $f_1(Y)$, $f_3(Y)$ and $g_3(Y)$ respectively have to be LD for Rank to be 2. ■

B.2 Proof of Lemma B.2

Proof. A multilinear monomial over $\{r_1, r_2, \ldots, r_n\}$ can be represented uniquely as an element of $\mathbb{F}_2^n$. For example, $r_2 r_3 r_{n-1}$ can be represented as the n-tuple $(0, 1, 1, 0, \ldots, 0, 1, 0)$. The multilinear monomials in a set are not mutually independent iff their corresponding set of n-tuples is linearly dependent over $\mathbb{F}_2$. If for a monomial $M \in S$ there are at most $\log|S| - 2$ mutually independent monomials in S, that are independent to M, then the n-tuples corresponding to all the monomials in S can be written as a linear combination of $\log|S| - 1$ tuples over $\mathbb{F}_2$. This is a contradiction as a set of $\log|S| - 1$ tuples can have at most $|S|/2$ distinct linear combinations over $\mathbb{F}_2$. Finally, as a multilinear monomial is a product of independent r.v.'s uniform over $\mathbb{F}^*$, it is easy to see that it is uniform over $\mathbb{F}^*$. ■

B.3 Proof of Irreducibility Lemma B.3

Proof. A multilinear polynomial $g(X)$(dependent on each x_i) is reducible iff there exists a proper partition $\{\{\bar{y}\}, \{\bar{z}\}\}$ of X such that $\text{Rank}_{YZ}(g) = 1$. The proof is by induction on $|X|$. In the base case if $|X| = \{x_1\}$ then, by $\texttt{SAMPLE}_2$, we have $f_R = a_1 x_1 + a_0$ where $a_0, a_1 \in_R \mathbb{F}$. Clearly f_R depends on x_1 with probability at least $1 - 1/|\mathbb{F}|$ and is irreducible. If $|X| = \{x_1, x_2\}$ then, by $\texttt{SAMPLE}_2$, we have $f_R = \alpha(a_1 x_1 + a_0)(b_1 x_2 + b_0) + \beta(c_1 x_1 + c_0)(d_1 x_2 + d_0)$ where independent $a_0, a_1, b_0, b_1, c_0, c_1, d_0, d_1, \alpha, \beta \in_R \mathbb{F}$. For $\text{Rank}_{\{x_1\}\{x_2\}}(f_R)$ to be 1, for all non-zero previous constants, it should be the case that $a_1.c_0 = a_0.c_1$ or $b_1.d_0 = b_0.d_1$. This holds with probability at most $10/|\mathbb{F}|$.

For $|X| = m$, let $f_R = \alpha A(\bar{a})B(\bar{b}) + \beta C(\bar{c})D(\bar{d})$ where $\{\{\bar{a}\}, \{\bar{b}\}\}$ and $\{\{\bar{c}\}, \{\bar{d}\}\}$ are partitions of X. As induction hypothesis we assume that A, B, C, D have Rank at least 2 w.r.t all proper partitions of their indicated variable sets. This also implies that they depend on every variable in their respective variable sets as if not, the partition of their variable set into ones which appear in the polynomial and the ones which do not would result in Rank 1 w.r.t this proper partition. Let $\{Y, Z\}$ be a proper partition of X. If either of $\text{Rank}_{YZ}(AB) > 1$ or $\text{Rank}_{YZ}(CD) > 1$ then, from Lemma 5.3, the probability that $\text{Rank}_{YZ}(f_R) = 1$ is at most $\frac{2^{O(m)}}{|\mathbb{F}|}$. So from now assume that Rank_{YZ} of both AB and CD is 1.

Note that if $\{\{\bar{a}\}, \{\bar{b}\}\} \neq \{Y, Z\}$ then both $\{\bar{a}\} \cap Y$ and $\{\bar{a}\} \cap Z$ are non-empty and thus $\{\{\bar{a}\} \cap Y, \{\bar{a}\} \cap Z\}$ is a non-trivial partition of $\{\bar{a}\}$. By induction hypothesis, this would imply that $\text{Rank}_{YZ}(A(\bar{a})) > 1$ which would further imply, from Lemma 5.3, that $\text{Rank}_{YZ}(f_R) = 1$ with probability at most $\frac{2^{O(m)}}{|\mathbb{F}|}$ (B vanishes with probability at most $1/|\mathbb{F}|$ as by construction it has a constant term). Hence, we can assume that $\{\{\bar{a}\}, \{\bar{b}\}\} = \{Y, Z\}$. Similarly, we can assume that $\{\{\bar{c}\}, \{\bar{d}\}\} = \{Y, Z\}$. Hence w.l.o.g we are left with the case when $f_R = \alpha A(Y)B(Z) + \beta C(Y)D(Z)$. Clearly, for $\text{Rank}_{YZ}(f_R)$ to be 1, for non-zero α, β, it should happen that $\exists \gamma_1, \gamma_2 \in \mathbb{F}$ s.t. $A(Y) = \gamma_1 C(Y)$ and $B(Z) = \gamma_2 D(Z)$. Note that at least one of $|Y|$ and $|Z|$ is at least 2. Let $|Y| \geq 2$. Then, by $\texttt{SAMPLE}_2$, with probability at least $1 - \frac{2^{O(m)}}{|\mathbb{F}|}$, $A(Y)$ is of the form $A = a_0 + a_1.M_Y + \ldots$ where M_Y is a monomial in Y variables, $a_0, a_1 \in_R \mathbb{F}$ and a_0, a_1 are indpendent. Let $C(Y) = c_0 + c_1.M_Y + \ldots$. For $A(Y) = \gamma_1 C(Y)$ and non-zero a_0, it should be the case that $a_1.c_0 = a_0.c_1$ and c_0 is non-zero. This holds with probability at most $2/|\mathbb{F}|$. ■

B.4 Proof of Lemma B.4

Proof. From the sampling method $\texttt{SAMPLE}_2(X, \mathbb{F})$ it is easy to see that, $f = \alpha.A(\bar{a})B(\bar{b}) + \beta.C(\bar{c})D(\bar{d})$ where $\{\{\bar{a}\}, \{\bar{b}\}\}$ and $\{\{\bar{c}\}, \{\bar{d}\}\}$ are randomly chosen partitions of X and $\alpha, \beta \in_R \mathbb{F}$. Also, A,B,C,D are

polynomials computed by multilinear formulas over their respective variable sets, sampled independently using $\texttt{SAMPLE}_2$ method. Hence, $|\{\bar{a}\} \cap Y| = Y_1 + Y_2 + \ldots + Y_{|Y|}$ where Y_i's are i.i.d 0-1 r.v.'s and $|\{\bar{b}\} \cap Y| = |Y| - |\{\bar{a}\} \cap Y|$. Using Chernoff's bound,

$$\Pr[\ |\{\bar{a}\} \cap Y| < |Y|/4 \text{ OR } |\{\bar{b}\} \cap Y| < |Y|/4\ ] \ \leq\ 2^{-|Y|/8}$$

Similarly, it follows that

$$\Pr[\ \min\{|\{\bar{a}\} \cap Y|, |\{\bar{a}\} \cap Z|\} < |Y|/4 \text{ OR } \min\{|\{\bar{b}\} \cap Y|, |\{\bar{b}\} \cap Z|\} < |Y|/4\ ] \ \leq\ 2^{-|Y|/9}$$

(1) Note that as A and B are multilinear polynomials on disjoint sets of variables, products of distinct pair of monomials in A and B are distinct monomials in AB. Hence, if both A and B have at least δ monomials in Z variables such that their coefficients are polynomials in Y variables each containing at least δ monomials then AB would have at least δ^2 such monomials in Z. Also, as $\alpha, \beta \in_R \mathbb{F}$ the probability, using union bound, that any of these monomials would be canceled by a monomial from CD is at most $\frac{2^m}{|\mathbb{F}|}$. In the worst case, $\min\{|\{\bar{a}\} \cap Y|, |\{\bar{a}\} \cap Z|\} = \min\{|\{\bar{b}\} \cap Y|, |\{\bar{b}\} \cap Z|\} = |Y|/4$. Now, applying induction and assuming the worst case each time we partition, we have the following expression where $\Delta(|Y' \cup Z'|, |Y'|)$ denotes the number of monomials (multilinear in Z') in a polynomial computed by a random formula over a set $Y' \cup Z'$ with $|Y'| = \min\{|Y'|, |Z'|\}$, such that their coefficients are polynomials in Y' each containing at least $\Delta(|Y' \cup Z'|, |Y'|)$ monomials.

$$\Delta(m, |Y|) \geq (\Delta(m/2, |Y|/4))^2 \ldots \geq \left(\Delta(m/2^h, |Y|/4^h)\right)^{2^h}.$$

For $2^h \leq \log \delta$, ensuring $|Y|/4^h \geq |Y|/\log^2 \delta$ and applying union bound on the failure probability each time we partition(we partition $O(m)$ times), we have with probability at least $1 - 2^{-\gamma m/18 \log^2 \delta} - \frac{2^{O(m)}}{|\mathbb{F}|}$, we have $\Delta(m, |Y|) \geq (\Delta(m/\log \delta, \gamma m/\log^2 \delta))^{\log \delta}$. Hence, all we need to show is that, $\Delta(m/\log \delta, \gamma m/\log^2 \delta) \geq 2$. This is easy to see as in the worst case, the formula sampled over variable set $Y \cup Z$ with both $|Y|, |Z| = \Omega(m)$ will be of the form $f' = \alpha.A'(Y)B'(Z) + \beta.C'(Y)D'(Z)$ where $\alpha, \beta \in_R \mathbb{F}$. Again, with the above stated probability one can show that there will be at least two monomials in each A',B',C',D' and the argument follows.

(2) Using the same argument from above we have, $f = \alpha.A(\bar{a})B(\bar{b}) + \beta.C(\bar{c})D(\bar{d})$. As from Lemma 5.3 with probability $1 - \frac{2^{O(m)}}{|\mathbb{F}|}$, $\text{Rank}_{YZ}(f) \geq \text{Rank}_{YZ}(AB) = \text{Rank}_{YZ}(A).\text{Rank}_{YZ}(B)$, we just need to show that $\text{Rank}_{YZ}(A) > 1$. Again with probability $1 - 2^{-\Omega(\gamma m)}$, A has the form $\alpha' A'(\bar{a}')B'(\bar{b}') + \beta' C'(\bar{c}')D'(\bar{d}')$ where for all $\bar{r}'$ there are $\Omega(m)$ elements of both Y and Z. From part (1) with probability $1 - 2^{-\Omega(\gamma m)} - \frac{2^{O(m)}}{|\mathbb{F}|}$ there are at least 2 monomials in each A',B',C',D' over Z variables such that there coefficients are polynomials over Y variables with at least 2 monomials. Hence, there is a monomial over Z variables in $A'B'$(and similarly in $C'D'$) such that its coefficient is a polynomial over Y variables with at least 4 monomials. If $\text{Rank}_{YZ}(A)$ is 1 then coefficients of these 2 monomials are multiples of each other. Let these coefficients be $h_1(Y)$ and $h_2(Y)$. Now as discussed above, coefficients of the monomials in $h_1(Y)$ have as their components multilinear monomials over a set of r.v.'s $\{r_1, \ldots, r_{|Y|}\}$. Similarly, coefficients of the monomials in $h_2(Y)$ have as their components multilinear monomials over a set of r.v.'s $\{s_1, \ldots, s_{|Y|}\}$. Hence with probability at least $1 - 1/|\mathbb{F}|$, $h_1(Y)$ and $h_2(Y)$ are LI and the lemma follows. ■