Fourier Growth of Parity Decision Trees

Uma Girish
Princeton University, NJ, USA

Avishay Tal
University of California at Berkeley, CA, USA

Kewen Wu
University of California at Berkeley, CA, USA

Abstract

We prove that for every parity decision tree of depth d on n variables, the sum of absolute values of Fourier coefficients at level ℓ is at most $d^{\ell/2} \cdot O(\ell \cdot \log(n))^{\ell}$. Our result is nearly tight for small values of ℓ and extends a previous Fourier bound for standard decision trees by Sherstov, Storozhenko, and Wu (STOC, 2021).

As an application of our Fourier bounds, using the results of Bansal and Sinha (STOC, 2021), we show that the k-fold Forrelation problem has (randomized) parity decision tree complexity $\widetilde{\Omega}\left(n^{1-1/k}\right)$, while having quantum query complexity $\lceil k/2 \rceil$.

Our proof follows a random-walk approach, analyzing the contribution of a random path in the decision tree to the level-ℓ Fourier expression. To carry the argument, we apply a careful cleanup procedure to the parity decision tree, ensuring that the value of the random walk is bounded with high probability. We observe that step sizes for the level-ℓ walks can be computed by the intermediate values of level $\le \ell - 1$ walks, which calls for an inductive argument. Our approach differs from previous proofs of Tal (FOCS, 2020) and Sherstov, Storozhenko, and Wu (STOC, 2021) that relied on decompositions of the tree. In particular, for the special case of standard decision trees we view our proof as slightly simpler and more intuitive.

In addition, we prove a similar bound for noisy decision trees of cost at most d – a model that was recently introduced by Ben-David and Blais (FOCS, 2020).

2012 ACM Subject Classification Theory of computation → Oracles and decision trees; Theory of computation → Communication complexity; Theory of computation → Quantum complexity theory

Keywords and phrases Fourier analysis of Boolean functions, noisy decision tree, parity decision tree, query complexity

Digital Object Identifier 10.4230/LIPIcs.CCC.2021.39

Funding *Uma Girish*: Research supported by the Simons Collaboration on Algorithms and Geometry, by a Simons Investigator Award and by the National Science Foundation grant No. CCF-1714779.

Acknowledgements We thank anonymous reviewers for helpful comments.

1 Introduction

A common theme in the analysis of Boolean functions is proving structural results on classes of Boolean devices (e.g., decision trees, bounded-depth circuits) and then exploiting the structure to: (i) devise pseudorandom generators fooling these devices, (ii) prove lower bounds, showing that some explicit function cannot be computed by such Boolean devices of certain size, or (iii) design learning algorithms for the class of Boolean devices in either the membership-query model or the random-samples model. Such structural results can involve properties of the Fourier spectrum of Boolean functions associated with Boolean devices, like concentration on low-degree terms or concentration on a few terms (i.e., "approximate sparsity").

In this work, we investigate the Fourier spectrum of parity decision trees. A parity decision tree (PDT) is an extension of the standard decision tree model. A PDT is a binary tree where each internal node is marked by a linear function (modulo 2) on the input variables $(x_1, \ldots, x_n)$, with two outgoing edges marked with 0 and 1, and each leaf is marked with either 0 or 1. A PDT naturally describes a computational model: on input $x = (x_1, \ldots, x_n)$, start at the root and at each step query the linear function specified by the current node on the input x and continue on the edge marked with the value of the linear function evaluated on x. Finally, when reaching a leaf, output the value specified in the leaf. PDTs naturally generalize standard decision trees that can only query the value of a single input bit in each internal node.

PDTs were introduced in the seminal paper of Kushilevitz and Mansour [21]. Aligned with the aforementioned theme, Kushilevitz and Mansour proved a structural result for PDTs and used it to design learning algorithms for PDTs. They showed that every PDT of size s computing a Boolean function $f\colon \{0,1\}^n \to \{0,1\}$ has

$$L_1(f) \triangleq \sum_{S \subseteq [n]} \left|\widehat{f}(S)\right| \leq s,$$

where $\widehat{f}(S)$ are the Fourier coefficients of f (see Subsection 2.1 for a precise definition). Then, they gave a learning algorithm in the membership-query model, running in time $\mathsf{poly}(t, n)$ that can learn any function f with $L_1(f) \leq t$. Combining the two results together, they obtained a $\mathsf{poly}(s, n)$-time algorithm for learning PDTs of size s.

Parity decision trees were also studied in relation to communication complexity and the log-rank conjecture [26, 39, 40, 38, 35, 31, 13, 20, 18, 33, 23]. Suppose Alice gets input $x \in \{0,1\}^n$, Bob gets input $y \in \{0,1\}^n$ and they want to compute some function $f(x, y)$. When f is an XOR-function, namely $f(x, y) = g(x \oplus y)$ for some $g : \{0,1\}^n \to \{0,1\}$, then any PDT for g of depth d can be translated into a communication protocol for f at cost $2d$: Alice and Bob simply traverse the PDT together, both exchanging the parity of their part of the input to simulate each query in the PDT. With this view, parity decision trees can be thought of as special cases of communication protocols for XOR functions. A surprising result by Hatami, Hosseini, and Lovett [18], shows that this is not far from the optimal strategy for XOR functions. Namely, if the communication cost for computing f is c, then the parity decision tree complexity of g is at most $\mathsf{poly}(c)$. Due to this connection, the log-rank conjecture for XOR-functions reduces to the question of whether Boolean functions with at most s non-zero Fourier coefficients can be computed by PDTs of depth $\mathsf{polylog}(s)$ [26, 39]. The best known upper bound is that such functions can be computed by PDTs of depth $O(\sqrt{s})$ [38] (or even non-adaptive PDTs of depth $\widetilde{O}(\sqrt{s})$ [33]).

While having small $L_1(f)$ norm implies learning algorithms and also simple pseudorandom generators fooling f [27], this property can be quite restrictive. In particular, very simple functions (e.g., the Tribes function) have $L_1(f)$ exponential in n. Such examples motivated Reingold, Steinke, and Vadhan [32] to study a more refined notion measuring for a given level ℓ, the sum of absolute values of Fourier coefficients of sets S of size exactly ℓ, i.e, to study

$$L_{1,\ell}(f) \triangleq \sum_{S \subseteq [n]: |S| = \ell} \left|\widehat{f}(S)\right|.$$

In particular, for $\ell = 1$, the measure $L_{1,1}(f)$ is tightly related to the total influence of f (and equals to it if f is monotone). The idea behind this more refined notion is that Fourier coefficients of different levels behave differently under standard manipulations to the function like random restrictions or noise operators. For example, when applying a noise operator with parameter γ, level-ℓ coefficients are multiplied by γ^ℓ. This motivates to establish a bound of the form $L_{1,\ell}(f) \leq t^\ell$ for some parameter t and all $\ell = 1, \dots, n$. If f satisfies such a bound, we say that $f \in \mathcal{L}_1(t)$.[1]

Reingold, Steinke, and Vadhan [32] showed that for read-once permutation branching programs of width w, while $L_1(f)$ could be exponential in n (even for $w = 3$), it nevertheless holds that $L_{1,\ell}(f) \leq (2w^2)^\ell$ for all $\ell = 1, \dots, n$. Then, they constructed a pseudorandom generator that fools any class of read-once branching programs for which $f \in \mathcal{L}_1(t)$ using only $t \cdot \mathsf{polylog}(n)$ random bits. This result was significantly generalized to a pseudorandom generator that fools any class of functions $f \in \mathcal{L}_1(t)$ using only $t^2 \cdot \mathsf{polylog}(n)$ random bits [9]. Further results established pseudorandom generators assuming $L_{1,\ell}$ bounds only on the first few levels [11, 8].

It turns out that read-once permutation branching programs are just one example of many well-studied Boolean devices with non-trivial $L_{1,\ell}$ bounds. The following classes of Boolean functions are other examples:

1. Width-w CNF and width-w DNF formulae are in $\mathcal{L}_1(O(w))$ [24].
2. AC^0 circuits of size s and depth d are in $\mathcal{L}_1\left(O(\log(s))^{d-1}\right)$ [36].
3. Boolean functions with max-sensitivity at most s are in $\mathcal{L}_1(O(s))$ [17]
4. Read-once branching programs of width w are in $\mathcal{L}_1\left(O(\log(n))^w\right)$[11]
5. Deterministic and randomized decision trees of depth d are in $\mathcal{L}_1\left(O\left(\sqrt{d\log(n)}\right)\right)$ [37, 34].
6. If $f(x,y)$ is a function computed by communication protocol exchanging at most c bits, then $h(z) = \mathbb{E}_x[f(x, x \oplus z)]$ satisfies $h \in \mathcal{L}_1(O(c))$ [15, 16].
7. Polynomials f over $\mathsf{GF}(2)$ of degree d have $L_{1,\ell}(f) \leq \left(2^{3d} \cdot \ell\right)^\ell$ [9].
8. Product tests, i.e., the XOR of multiple Boolean functions operating on disjoint sets of at most m bits each, are in $\mathcal{L}_1(O(m))$ [22].

We remark that Items 1, 2, 4, 5 and 8 are essentially tight, Item 3 can be potentially improved polynomially [28, 30], Item 6 can be potentially improved quadratically [15] and Item 7 can be potentially improved exponentially [10]. Indeed, improving Item 7 exponentially would imply that $\mathsf{AC}^0[\oplus]$ in $\mathcal{L}_1(\mathsf{polylog}(n))$ and would give the first poly-logarithmic pseudorandom generators for this well-studied class of Boolean circuits [10].

The most relevant result to our work is the recent tight bounds on the $L_{1,\ell}$ of decision trees of depth d. Sherstov, Storozhenko and Wu [34] recently proved that for any randomized decision tree of depth d computing a function f, it holds that $L_{1,\ell}(f) \leq \sqrt{\binom{d}{\ell} \cdot O(\log(n))^{\ell-1}}$. Their bound is nearly tight (see [37, Section 7] and [29, Chapter 5.3] for tightness examples). One motivation for showing such a bound for decision trees is that it demonstrates a stark difference between quantum algorithms making few queries and randomized algorithms making a few queries. Indeed, the Fourier spectrum associated with quantum query algorithms making a few queries can be far from being approximately sparse (in the sense that its $L_{1,\ell}$ is quite large). Based on that difference, both [34] and [2] showed that there are partial functions, either k-fold Forrelation or k-fold Rorrelation, that can be correctly computed with probability at least $1/2 + \Omega(1)$ by quantum algorithms making $\lceil k/2 \rceil$ queries, but require $\widetilde{\Omega}\left(n^{1-1/k}\right)$ queries for any randomized algorithm. Moreover, due to the result of Aaronson and Ambainis [1] this is the largest possible separation between the two models.

[1] Note that if $f \in \mathcal{L}_1(t)$ then after applying noise operator with $\gamma = 1/(2t)$, the noisy-version of f has total L_1-norm at most $O(1)$ which makes it is quite easy to fool using small-biased distributions [27].

Indeed, as suggested in [37], one can show that any function with sufficiently good bounds on its $L_{1,\ell}$, for all $\ell = 1, \dots, n$, cannot solve the k-fold Rorrelation, and such bounds were obtained by [34] for randomized decision trees of depth $n^{1-1/k}/\mathsf{polylog}(n)$. Independently, Bansal and Sinha obtained the same separation but only relying on the $L_{1,\ell}$ bounds for $\ell \in \{k, k+1, \dots, k^2\}$. With this additional flexibility, they were able to obtain their separation for the simpler and explicit function called k-fold Forrelation.

For parity decision trees, the work of Blais, Tan, and Wan [4] established a tight bound of $O\left(\sqrt{d}\right)$ on the first level $\ell = 1$. To the best of our knowledge, bounds on higher levels were not considered previously in the literature (in fact, even for standard decision trees, such bounds were not considered prior to [37]).

1.1 Our Results

We prove level-ℓ bounds for any parity decision tree of depth d.

▶ **Theorem 1** (Informal). *Let $\mathcal{T}$ be a depth-d parity decision tree on n variables. Then the sum of absolute Fourier coefficients at level ℓ is bounded by $d^{\ell/2} \cdot O(\ell \cdot \log(n))^{\ell}$.*

See Theorem 32 and Theorem 39 for a precise statement taking into account the probability that $\mathcal{T}$ accepts a uniformly random input. Theorem 1 extends the result of [34] from standard decision trees to parity decision trees at the cost of an $(\ell \cdot \log(n))^{O(\ell)}$ multiplicative factor. We remark that even for standard decision tree there is a lower bound of $L_{1,\ell}(f) \geq \sqrt{\binom{d}{\ell} \cdot (\log(n))^{\ell-1}}$ [37, Section 7] for constant ℓ and $L_{1,\ell}(f) \geq \frac{1}{\mathsf{poly}(\ell)} \cdot \sqrt{\binom{d}{\ell}}$ for all ℓ [29, Chapter 5.3]. Thus, our bounds are tight up to $\mathsf{polylog}(n)$ factors for constant ℓ, and they deteriorate as ℓ grows. Nevertheless, our main application relies on the bounds for small values of ℓ (constant or at most $\log^2 n$).

Noisy Decision Trees

We also investigate the Fourier spectrum of noisy decision trees. Noisy decision trees are a different generalization of the standard model; here in each internal node v we query a noisy version of an input bit, that equals the true bit with probability $(1+\gamma_v)/2$. Any such query costs γ_v^2. We say that a noisy decision tree has cost at most d if the total cost in any root-to-leaf path is at most d. Recent work studied this model and established connections to the question of how randomized decision tree complexity behaves under composition [3].

We prove level-ℓ bounds for any noisy decision tree of cost at most d. See Theorem 42 for a precise statement.

▶ **Theorem 2** (Informal). *Let $\mathcal{T}$ be a noisy decision tree of cost at most d on n variables. Then the sum of absolute Fourier coefficients at level ℓ is bounded by $O(d)^{\ell/2} \cdot (\ell \cdot \log(n))^{(\ell-1)/2}$.*

Extension to Randomized Query Models

It is simple to verify that if f is a convex combination of Boolean functions $f_1, \dots, f_m$ each with $L_{1,\ell}(f_i) \leq t_\ell$ then also f satisfy $L_{1,\ell}(f) \leq t_\ell$. Thus, if we take a distribution over PDTs of depth d (resp., noisy decision trees of cost d) we get the same bounds on their $L_{1,\ell}$ as those in Theorem 1 (resp., Theorem 2). This is captured in the following corollary.

▶ **Corollary 3.** *Let $\mathcal{T}$ be a randomized parity decision tree of depth at most d on n variables. Then,*

$$\forall \ell \in [n] : L_{1,\ell}(\mathcal{T}) \le d^{\ell/2} \cdot O(\ell \cdot \log(n))^{\ell}.$$

Let $\mathcal{T}'$ be a randomized noisy decision tree of cost at most d on n variables. Then,

$$\forall \ell \in [n] : L_{1,\ell}(\mathcal{T}') \le O(d)^{\ell/2} \cdot (\ell \cdot \log(n))^{(\ell-1)/2}.$$

1.2 Applications

Quantum versus Randomized Query Complexity

Let $k \le \log(n)$. Bansal and Sinha [2] gave a $\lceil k/2 \rceil$ versus $\widetilde{\Omega}\left(n^{1-1/k}\right)$ separation between the quantum and randomized query complexity of k-fold Forrelation (defined by [1]). For our purposes just think of k-fold Forrelation as a partial Boolean function on n input bits. Our main application is an extension of Bansal and Sinha's lower bound for the model of randomized parity decision trees. This follows from their main technical result and Theorem 1.

▶ **Theorem 4** (Restatement of [2, Theorem 3.2])**.** *Let $f\colon \{0,1\}^n \to [0,1]$ such that f and all its restrictions satisfy $L_{1,\ell}(f) \le t^{\ell}$ for $\ell = \{k, \dots, k(k-1)\}$. Let $\delta = 2^{-5k}$. Suppose f is δ-close to the value of k-fold Forrelation of x for all x on which k-fold Forrelation is defined. Then, $t \ge \Omega\left(\frac{n^{(1-1/k)/2}}{k^{15}}\right)$.*

▶ **Corollary 5.** *If $\mathcal{T}$ is a randomized parity decision tree of depth d computing k-fold Forrelation with success probability $\frac{1}{2} + \gamma$, then $d \ge \gamma^2 \cdot \frac{n^{1-1/k}}{\mathsf{poly}(k)\log^2 n}$.*

Proof. We can amplify the success probability of the randomized parity decision tree from $1/2+\gamma$ to $1-2^{-5k}$ by repeating the query algorithm $O(k/\gamma^2)$ times independently and taking majority. This results in a randomized parity decision tree $\mathcal{T}'$ of depth $d' = O(d \cdot k/\gamma^2)$. Now, Corollary 3 gives $L_{1,\ell}(\mathcal{T}') \le (d')^{\ell/2} \cdot O(\ell \cdot \log(n))^{\ell}$ for all ℓ. In particular, $L_{1,\ell}(\mathcal{T}') \le t^{\ell}$ for all $\ell \le k(k-1)$ where $t = O\left(\sqrt{d'} \cdot k(k-1) \cdot \log(n)\right)$. This is also true for any restriction of $\mathcal{T}'$, since fixing variables to constants yields another randomized parity decision tree of depth at most d'. Combining the bounds on $L_{1,\ell}(\mathcal{T}')$ for $\ell \in \{k, \dots, k(k-1)\}$ with Theorem 4 gives $d' \ge \frac{n^{1-1/k}}{O(k^{34}) \cdot \log^2(n)}$ and thus $d \ge \gamma^2 \cdot \frac{n^{1-1/k}}{O(k^{35}) \cdot \log^2(n)}$. ◀

For constant k and $\gamma = 2^{-O(k)}$, we get a $\lceil k/2 \rceil$ versus $\widetilde{\Omega}\left(n^{1-1/k}\right)$ separation between the quantum query complexity and the randomized parity query complexity of k-fold Forrelation. We remark that separations in the reverse direction are also known: for the n-bit parity function, the (randomized) parity query complexity is 1 whereas the quantum query complexity is $\Omega(n)$ [25].

Similarly, we can obtain the following corollary for noisy decision trees.

▶ **Corollary 6.** *If $\mathcal{T}$ is a randomized noisy decision tree of cost at most d computing k-fold Forrelation with success probability $\frac{1}{2} + \gamma$, then $d \ge \gamma^2 \cdot \frac{n^{1-1/k}}{\mathsf{poly}(k)\log(n)}$.*

Towards Communication Complexity Lower Bounds

We recall an open question from [15], which, if true, would demonstrate that the randomized communication complexity of the Forrelation problem composed with the XOR gadget is $\widetilde{\Omega}(n^{1/2})$. The *simultaneous* quantum communication complexity of this problem is $\mathsf{polylog}(n)$ and the best known randomized lower bound is $\widetilde{\Omega}(n^{1/4})$ due to [15].

▶ **Conjecture 7.** *Let $f\colon \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ computed by a deterministic communication protocol of cost at most c. Let $h : \{0,1\}^n \to [0,1]$ defined by $h(z) = \mathbb{E}_x[f(x, x \oplus z)]$. Then, $L_{1,2}(h) \leq c \cdot \mathsf{polylog}(n)$.*

We view Theorem 1 as a first step towards this conjecture. Indeed, for communication protocols that follow a parity decision tree strategy according to some tree $\mathcal{T}$, it is simple to verify that $h = \mathcal{T}$ (as functions), and thus $L_{1,2}(h) = L_{1,2}(\mathcal{T}) \leq c \cdot \mathsf{polylog}(n)$.

We remark that there is a separation of $\mathsf{polylog}(n)$ versus $\widetilde{\Omega}(n^{1/2})$ between *simultaneous* quantum communication complexity and *two-way* randomized communication complexity due to [14]. We also know a separation of $O(k \log n)$ versus $\widetilde{\Omega}(n^{1-1/k})$ between *two-way* quantum communication complexity and *two-way* randomized communication complexity. This can be obtained by combining the optimal quantum versus classical query complexity separations of [2] and [34] and the query-to-communication lifting theorems [7] using the inner product gadget.

Application to Expander Random Walk

Recently, [12] showed that expander random walks fool symmetric functions and also general functions in $\mathcal{L}_1(t)$. To be more precise, assume $f \in \mathcal{L}_1(t)$. Let G be an expander, with second eigenvalue $\lambda \ll \frac{1}{t^4}$, where half of G's vertices are labeled by 0 and the rest are labeled by 1. Then the expected value of f on bits sampled by an $(m-1)$-step random walk on G is approximately the value it would get on a uniformly random string in $\{0,1\}^m$. Combined with our results, this shows that if f can be computed by low-depth parity decision trees then f can be fooled by the expander random walk.

Fourier Bounds for Small-size Parity Decision Trees

By a simple size-to-depth reduction we obtain Fourier bounds for parity decision trees of bounded size. We defer the simple proof to Appendix A.

▶ **Corollary 8.** *Let $\mathcal{T}$ be a parity decision tree of size at most $s > 1$ on n variables. Then,*

$$\forall \ell \in [n] : L_{1,\ell}(f) \leq (\log(s))^{\ell/2} \cdot O(\ell \cdot \log(n))^{1.5\ell}.$$

1.3 Technical Overview

For the rest of the paper we consider Boolean functions as functions from $\{\pm 1\}^n$ to $\{0,1\}$. This is for convenience, since most of our calculations become easier under this representation. Observe that under this view, a parity decision tree queries at each internal node the product $\prod_{i \in S} x_i$ for some $S \subseteq [n]$ and goes left/right depending on whether $\prod_{i \in S} x_i = 1$ or -1.

Let $\ell \in \mathbb{N}_+$. For simplicity of notation, we use $\widetilde{O}_\varepsilon(d^m)$ to denote $\left(d \cdot \mathsf{polylog}\left(n^\ell/\varepsilon\right)\right)^m$ for $m, n, d \in \mathbb{N}_+$ and $\varepsilon \in (0, 1/2]$. When we omit the subscript ε, it is understood that $\varepsilon = 1$. As per this notation, we show a bound of $\widetilde{O}\left(d^{\ell/2}\right)$ on the level-ℓ Fourier mass of parity decision trees of depth d. We first describe the proof for standard decision trees and then show how to generalize to parity decision trees.

Standard Decision Trees

Let $\mathcal{T}$ be a decision tree and for simplicity, assume that every leaf is of depth d. Let $v_0, \ldots, v_d$ be a random root-to-leaf path in $\mathcal{T}$ and $\boldsymbol{v}^{(0)}, \ldots, \boldsymbol{v}^{(d)} \in \{-1, 0, 1\}^n$ denote the sequence of partial assignments, i.e., for $j \in [n]$ and $i \in \{0, \ldots, d\}$, let

$$\boldsymbol{v}_j^{(i)} = \begin{cases} 1 & \text{if } x_j \text{ is fixed to } 1 \text{ before reaching } v_i, \\ -1 & \text{if } x_j \text{ is fixed to } -1 \text{ before reaching } v_i, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

For $u \in \mathbb{R}^n$, we use u_S to denote $\prod_{j\in S} u_j$. Let $a_S = \mathsf{sgn}\left(\widehat{\mathcal{T}}(S)\right)$ for $|S| = \ell$ and 0 otherwise. Note that

$$\sum_{S:|S|=\ell} \left|\widehat{\mathcal{T}}(S)\right| = \sum_{S:|S|=\ell} a_S \widehat{\mathcal{T}}(S) = \sum_{S:|S|=\ell} a_S \mathop{\mathbb{E}}_{v_d}\left[\mathcal{T}(v_d)\boldsymbol{v}_S^{(d)}\right] = \mathop{\mathbb{E}}_{v_d}\left[\mathcal{T}(v_d)\sum_{S:|S|=\ell} a_S \boldsymbol{v}_S^{(d)}\right]. \tag{2}$$

Thus, to bound $\sum_{S:|S|=\ell} |\widehat{\mathcal{T}}(S)|$ it suffices to show that $\left|\sum_{S:|S|=\ell} a_S \cdot \boldsymbol{v}_S^{(d)}\right|$ is bounded by $\widetilde{O}(d^{\ell/2})$ in expectation. Denote by $X^{(i)} := \sum_{S:|S|=\ell} a_S \cdot \boldsymbol{v}_S^{(i)}$ for $i = 0, 1, \ldots, d$. We write $X^{(d)}$ as a telescoping sum $X^{(d)} = \sum_{i=1}^d \left(X^{(i)} - X^{(i-1)}\right)$. To analyze the difference sequence, observe that in the expression

$$X^{(i)} - X^{(i-1)} = \sum_{S:|S|=\ell} a_S \cdot \left(\boldsymbol{v}_S^{(i)} - \boldsymbol{v}_S^{(i-1)}\right),$$

if set S contributes to the sum, then S must include the bit queried at the $(i-1)$-th step of the path. Conditioning on $v_0, \ldots, v_{i-1}$, let x_j be the variable queried in v_{i-1}, then we have

$$X^{(i)} - X^{(i-1)} = \sum_{S:|S|=\ell, j\in S} a_S \cdot \boldsymbol{v}_S^{(i)} \; = \; x_j \cdot \left(\sum_{S:|S|=\ell, j\in S} a_S \cdot \boldsymbol{v}_{S\setminus\{j\}}^{(i-1)}\right).$$

Furthermore, we observe that the sum $\sum_{S:|S|=\ell, j\in S} a_S \cdot \boldsymbol{v}_{S\setminus\{j\}}^{(i-1)}$ is determined by v_{i-1}; thus conditioning on $v_0, \ldots, v_{i-1}$ the value of $X^{(i)} - X^{(i-1)}$ is a random coin in $\{\pm 1\}$ multiplied by some fixed integer. In other words, we get that $X^{(0)}, \ldots, X^{(d)}$ is a martingale with varying step sizes.

Recall that Azuma's inequality provides concentration bounds for martingales with bounded step sizes, thus now we need to bound $\left|\sum_{S:|S|=\ell, j\in S} a_S \cdot \boldsymbol{v}_{S\setminus\{j\}}^{(i-1)}\right|$, which is similar to our initial goal. Put differently, we wish to analyze the sum

$$\sum_{S'\subseteq[n]\setminus\{j\}:|S'|=\ell-1} a_{S'\cup\{j\}} \cdot \boldsymbol{v}_{S'}^{(i-1)},$$

which calls for an inductive argument on ℓ. In addition, since we eventually apply a union bound on all steps, we need to show that $\left|\sum_{S'} a_{S'\cup\{j\}} \boldsymbol{v}_{S'}^{(i-1)}\right|$ is bounded with high probability (and not just in expectation).

More generally, to carry an inductive argument we define for any set $T \subseteq [n]$, $|T| \le \ell$ and any $i \in \{0, \ldots, d\}$, the random variable

$$X_T^{(i)} := \sum_{S\supseteq T:|S|=\ell} a_S \cdot \boldsymbol{v}_{S\setminus T}^{(i)} = \sum_{S'\subseteq\overline{T}:|S'|=\ell-|T|} a_{S'\cup T} \cdot \boldsymbol{v}_{S'}^{(i)}.$$

Note that our initial goal was to bound $\left|X_\emptyset^{(d)}\right| = \left|X^{(d)}\right|$, which is analyzed by (reverse) induction on $|T|$ going from larger sets to smaller sets as Lemma 9.

▶ **Lemma 9.** *For all $t \in \{0, \ldots, \ell\}$ and $\varepsilon > 0$, the probability that there exist $i \in \{0, \ldots, d\}$ and $T \subseteq [n]$ of size at least t such that $\left|X_T^{(i)}\right| \geq \widetilde{O}_\varepsilon\left(d^{(\ell-t)/2}\right)$ is at most $\varepsilon \cdot (\ell - t)$.*

The main observation for the proof is that $X_T^{(0)}, X_T^{(1)}, \ldots, X_T^{(d)}$ is a martingale whose difference sequence consists of terms of the form $X_{T'}^{(i-1)}$ where $T \subsetneq T'$. To see this, if we are querying x_j at v_{i-1}, then

$$X_T^{(i)} - X_T^{(i-1)} = \begin{cases} 0 & j \in T, \\ x_j \cdot \left(\sum_{j \notin S \subseteq \overline{T}} a_{S \cup T \cup \{j\}} \cdot \boldsymbol{v}_S^{(i-1)} \right) = x_j \cdot X_{T \cup j}^{(i-1)} & j \notin T. \end{cases}$$

Note that $X_{T\cup j}^{(i-1)}$ depends only on the history until v_{i-1}, and x_j is a uniformly random bit independent of this history, thus $X_T^{(i)}$ is a martingale. The inductive hypothesis implies that with at least $1 - \varepsilon \cdot (\ell - t - 1)$ probability, $\left|X_{T\cup j}^{(i-1)}\right| \leq \widetilde{O}_\varepsilon\left(d^{(\ell-t-1)/2}\right)$ for all T of size t and $j \in [n] \setminus T$. Whenever this happens, Azuma's inequality implies that[2] with probability at least $1 - \varepsilon / \left(d \cdot n^t\right)$, we have

$$\left|X_T^{(i)}\right| \leq 2\sqrt{\log(d \cdot n^t/\varepsilon)} \cdot \sqrt{\sum_{i=1}^{d} \widetilde{O}_\varepsilon\left(d^{\ell-t-1}\right)} = \widetilde{O}_\varepsilon\left(d^{(\ell-t)/2}\right).$$

This, along with a union bound over T of size t and $i \in \{0, \ldots, d\}$ completes the inductive step. The Fourier bound for noisy decision trees can be proved using a similar approach.

Parity Decision Trees

The basic approach is as before. Let $\mathcal{T}$ be a parity decision tree. As in (1), we use v_i and $\boldsymbol{v}^{(i)}$ to denote the random walk and the partial assignments to the variables respectively. We say v_i is *k-clean* if

$$\forall S \subseteq [n], |S| \leq k, \quad \boldsymbol{v}_S^{(i)} = \begin{cases} 1 & \text{if } x_S \text{ is fixed to 1 before reaching } v_i, \\ -1 & \text{if } x_S \text{ is fixed to } -1 \text{ before reaching } v_i, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

For (2) to be true, we need that at least v_d is ℓ-clean. Note that this is not always true,[3] but it is useful as it simplifies the study of high-level Fourier coefficients. To address this issue, we define a *cleanup* process for parity decision trees in which we make additional queries to ensure that certain key nodes are k-clean. We do this by recursively cleaning nodes in a top-down fashion so that for every node v in the original tree $\mathcal{T}$, any node v' in the new tree $\mathcal{T}'$ obtained at the end of the cleanup step for v is k-clean.

The cleanup process is simple to describe: Let $v_1, \ldots, v_d$ be any root-to-leaf path in $\mathcal{T}$. Assume we have completed the cleanup process for $v_1, \ldots, v_{i-1}$. We then query the parity at v_i. While there exists a (minimal) set S violating (3), we pick and query an arbitrary

[2] Technically this is not true, since a martingale after conditioning may not still be a martingale. We handle this by truncating the martingale when a bad event happens instead of conditioning on the good event.

[3] For example, let $S = \{1, 2\}$ and consider the parity decision tree whose only query is x_1x_2. At any leaf, the value of x_1x_2 is fixed, however, the values of x_1 and x_2 are free, hence S violates (3).

coordinate in S. Once (3) is satisfied, we proceed to the cleanup process for v_{i+1}. This process increases the depth by a factor of at most k. We set $k = \Theta(\ell \cdot \log(n))$ and work with the new tree $\mathcal{T}'$ of depth $D \leq k \cdot d$.

Let $v_0, \dots, v_D$ be a random root-to-leaf path in $\mathcal{T}'$ and $I_i, i \in [D]$ be the set of coordinates fixed due to the query at v_{i-1}. Note that this set might be of size larger than 1.[4] It follows from simple linear algebra that $\sum_{i=1}^{D} |I_i| \leq D$. Since v_D is k-clean, (2) holds. Defining $X_T^{(i)}$ exactly as before, our goal is to prove Lemma 9 with D instead of d. The proof is still by induction on $\ell - t$. It turns out that $X_T^{(0)}, X_T^{(1)}, \dots, X_T^{(D)}$ is no longer a martingale; instead, $X_T^{(i)} - X_T^{(i-1)} = Y_i + Z_i$ where

$$Y_i := \sum_{\substack{\emptyset \neq J \subseteq I_i \cap \overline{T} \\ |J| \text{ is even}}} x_J \cdot X_{J \cup T}^{(i-1)} \quad \text{and} \quad Z_i := \sum_{\substack{\emptyset \neq J \subseteq I_i \cap \overline{T} \\ |J| \text{ is odd}}} x_J \cdot X_{J \cup T}^{(i-1)}. \tag{4}$$

and Z_i (resp., Y_i) is an odd (resp., even) polynomial of degree at most ℓ over the newly fixed variables $\{x_j \mid j \in I_i\}$. Conditioning on v_{i-1}, every pair of random bits $(x_j, x_{j'})$ from $\{x_j \mid j \in I_i\}$ is either identical $(x_j \equiv x_{j'})$ or opposite $(x_j \equiv -x_{j'})$, which means Y_i is a constant and Z_i can be written as $z_i \cdot |Z_i|$ where $|Z_i|$ is a constant and $z_i \sim \{\pm 1\}$.

For now, let us ignore Y_i and assume that we have a martingale $X_T^{(i)}$ such that $X_T^{(i)} - X_T^{(i-1)} = z_i \cdot |Z_i|$, where $z_i \sim \{\pm 1\}$ is a uniformly random bit independent of $z_0, \dots, z_{i-1}$ and $|Z_i|$ depends only on v_{i-1}. Combined with an adaptive version of Azuma's inequality, we only need to show the sum of squares of step sizes $\sum_{i=1}^{D} |Z_i|^2$ is $\widetilde{O}_\varepsilon\left(D^{\ell-t}\right)$ to prove $\left|X_T^{(i)}\right| = \widetilde{O}_\varepsilon\left(D^{(\ell-t)/2}\right)$. By the induction hypothesis, with probability at least $1 - \varepsilon \cdot (\ell - t - 1)$ the coefficients of Z_i are bounded appropriately. Since $\sum_{i=1}^{D} |I_i| \leq D$ and in particular $|I_i| \leq D$, we have

$$|Z_i| \leq \sum_{\text{odd } j \geq 1} \binom{|I_i|}{j} \cdot \max_{|T'|=j+t} \left|X_{T'}^{(i-1)}\right| \leq \sum_{j \geq 1}^{\ell-t} \binom{|I_i|}{j} \cdot \widetilde{O}_\varepsilon\left(D^{(\ell-j-t)/2}\right) = \widetilde{O}_\varepsilon\left(|I_i| \cdot D^{(\ell-t-1)/2}\right)$$

and thus $\sum_{i=1}^{D} |Z_i|^2 \leq D^2 \cdot \widetilde{O}_\varepsilon\left(D^{\ell-t-1}\right)$. This is too loose for our purpose.

We instead try to bound the sum of squares of step sizes *with high probability*. Imagine for now that v_{i-1} is 2-clean.[5] Then, the variables $\{x_j \mid j \in I_i\}$ are 2-wise independent conditioning on v_{i-1}. This gives

$$\begin{aligned} \mathbb{E}\left[|Z_i|^2 \,\middle|\, v_{i-1}\right] &\leq \sum_{\text{odd } j \geq 1} \binom{|I_i|}{j} \cdot \max_{|T'|=j+t} \left|X_{T'}^{(i-1)}\right|^2 \\ &\leq \sum_{j \geq 1}^{\ell-t} \binom{|I_i|}{j} \cdot \widetilde{O}_\varepsilon\left(D^{\ell-j-t}\right) = \widetilde{O}_\varepsilon\left(|I_i| \cdot D^{\ell-t-1}\right) \end{aligned}$$

and thus $\mathbb{E}\left[\sum_{i=1}^{D} |Z_i|^2\right] \leq \widetilde{O}_\varepsilon\left(D^{\ell-t}\right)$. To show this bound holds with high probability, we use concentration properties of degree-ℓ polynomials under k-wise independent distributions for $k \gg \ell$.

[4] For example, suppose we query x_1x_2, x_1x_3, x_1x_4 and finally x_1. Then, the last query reveals 4 coordinates.

[5] This assumption immediately implies that $|I_i| \leq 1$ and trivially proves our inequality, however, this type of reasoning doesn't generalize to the case when v_{i-1} is not 2-clean.

In the actual proof, we proceed by conditioning on $C(v_{i-1})$, the nearest ancestor of v_{i-1} that is k-clean, instead of conditioning on v_{i-1}, which allows to remove the assumption that v_{i-1} is 2-clean. This is because the queries within a cleanup step are non-adaptive, thus Z_i depends only on $C(v_{i-1})$ and not on v_{i-1}.

Meanwhile, although $X_T^{(i)}$ is not quite a martingale sequence (due to Y_i) and the step sizes (i.e., $|Z_i|$) are adaptive and not always bounded, we are nonetheless able to prove an adaptive version of Azuma's inequality of the form $\mathbf{Pr}\left[\max_{i\in[D]}\left|X_T^{(i)}\right| \geq \mu + t\cdot\sigma\right] \leq e^{-\Omega(t^2)} + \varepsilon$ provided $\mathbf{Pr}\left[\left(\sum_{i=1}^{D}|Y_i| \leq \mu\right) \wedge \left(\sum_{i=1}^{D}|Z_i|^2 \leq \sigma^2\right)\right] \geq 1-\varepsilon$. Then it suffices to bound $\sum_{i=1}^{D}|Y_i|$ similarly to $\sum_{i=1}^{D}|Z_i|^2$ above.

1.4 Related Work

We remark that our proof for level-ℓ Fourier growth (even when specialized to the case of standard decision trees) differs from the proofs appearing in [37] and [34]. There, the results were based on decompositions of decision trees. We view our martingale approach as natural and intuitive. We wonder if one can obtain the tight results from [34] using this approach. It seems that the main bottleneck is a union bound on events related to all sets $T \subseteq [n]$ of size at most ℓ.

Our bounds for level-1 improve those obtained by [4]. They prove that $L_{1,1}(\mathcal{T}) \leq O(\sqrt{p\cdot d})$ when $p = \mathbf{Pr}_x[\mathcal{T}(x) = 1]$, whereas we obtain a bound of

$$L_{1,1}(\mathcal{T}) \leq O\left(p\sqrt{d}\cdot\log(1/p)\right).$$

In particular, our bound is almost quadratically better for small values of p. It remains open whether the bound can be further improved to $O\left(p\sqrt{d\cdot\log(1/p)}\right)$, which is the optimal bound for standard decision trees.

We remark that our cleanup technique is inspired by [4], which used cleanup to prove their level-1 bound. However, our proof strategies and the way we use the cleanup procedure is quite different than that of [4].

Organization

We make formal definitions in Section 2. We state and prove the necessary concentration inequalities in Section 3. We present the cleanup process in Section 4. We present the Fourier bounds for parity decision trees in Section 5 and for noisy decision trees in Section 6.

2 Preliminaries

We use $\log(\cdot)$ to denote the logarithm with base 2. We use $[n]$ to denote $\{1,2,\ldots,n\}$; and $\binom{[n]}{k}$ (resp., $\binom{[n]}{\leq k}$) to denote the set of all size-k (resp., size-at-most-k) sets from $[n]$. If S is a set from universe U, then we write $\overline{S}$ for $U \setminus S$. We use $\mathcal{U}_n$ to denote the uniform distribution over $\{\pm 1\}^n$. We use $\mathsf{sgn}(\mathsf{value}) \in \{-1,0,1\}$ to denote the sign of value, i.e., $\mathsf{sgn}(\mathsf{value})$ equals -1 if $\mathsf{value} < 0$, 1 if $\mathsf{value} > 0$, and 0 if $\mathsf{value} = 0$.

We use $\mathbb{F}_2 = \{0,1\}$ to denote the binary field, $\mathsf{Span}\langle\mathsf{vectors}\rangle$ to denote the subspace spanned by $\mathsf{vectors}$ over $\mathbb{F}_2$. For a distribution $\mathcal{D}$ we use $x \sim \mathcal{D}$ to represent that x is a random variable sampled from $\mathcal{D}$. For a finite set $\mathcal{X}$ we use $x \sim \mathcal{X}$ to denote that x is a random variable sampled uniformly from $\mathcal{X}$. We use the standard notion of k-wise independent distribution over $\{\pm 1\}^n$.

▶ **Definition 10** (k-wise independence). *A distribution $\mathcal{D}$ over $\{\pm 1\}^n$ is k-wise independent if for $x \sim \mathcal{D}$ and any k-indices $1 \le i_1 < i_2 < \ldots < i_k \le n$, the random variables $(x_{i_1}, \ldots, x_{i_k})$ are uniformly distributed over $\{\pm 1\}^k$.*

2.1 Boolean Functions

Here we recall definitions in the analysis of Boolean functions (see [29] for a detailed introduction). Let $f\colon \{\pm 1\}^n \to \mathbb{R}$ be any Boolean function. For any $p > 0$, the p-norm of f is defined as $\|f\|_p = \left(\mathbb{E}_{x\sim\mathcal{U}_n}\left[|f(x)|^p\right]\right)^{1/p}$. For any subset $S \subseteq [n]$, x_S denotes $\prod_{i\in S} x_i$ (in particular, $x_\emptyset = 1$). It is a well-known fact that we can uniquely represent f as a linear combination of $\{x_S\}_{S\subseteq[n]}$:

$$f(x) = \sum_{S\subseteq[n]} \widehat{f}(S) x_S,$$

where the coefficients $\left\{\widehat{f}(S)\right\}_{S\subseteq[n]}$ are referred to as the *Fourier coefficients* of f and are given by $\widehat{f}(S) = \mathbb{E}_{x\sim\mathcal{U}_n}\left[f(x)x_S\right]$. The above representation expresses f as a multilinear polynomial and is called the Fourier representation of f. We say that f is of degree at most d if its Fourier representation is a polynomial of degree at most d, i.e., if $\widehat{f}(S) = 0$ for all $S \subseteq [n], |S| > d$.

2.2 Parity Decision Trees

Here we formally define parity decision trees (with Boolean outputs).

▶ **Definition 11** (Parity decision tree). *A* parity decision tree $\mathcal{T}$ *is a representation of a Boolean function $f\colon \{\pm 1\}^n \to \{0,1\}$. It consists of a rooted binary tree in which each internal node v is labeled by a non-empty set $Q_v \subseteq [n]$, the outgoing edges of each internal node are labeled by $+1$ and -1, and the leaves are labeled by 0 and 1.*

On input $x \in \{\pm 1\}^n$, the tree $\mathcal{T}$ constructs a computation path *$\mathcal{P}$ from the root to a leaf. Specifically, when $\mathcal{P}$ reaches an internal node v we say that $\mathcal{T}$* queries *Q_v; then $\mathcal{P}$ follows the outgoing edge labeled by $\prod_{i\in Q_v} x_i$. We require that Q_v is not implied by its ancestors' queries. The output of $\mathcal{T}$ (and hence f) on input x is the label of the leaf reached by the computation path. Conversely, we say x is* consistent with *the path $\mathcal{P}$ if $\mathcal{P}$ is the computation path (possibly ending before reaching a leaf) for x.*

We make a few more remarks on a parity decision tree $\mathcal{T}\colon \{\pm 1\}^n \to \{0,1\}$.

- A node v in $\mathcal{T}$ can be either an internal node or a leaf, and we use $\mathcal{T}(v) \in \{0,1\}$ to denote the label on v when v is a leaf. Meanwhile, we use $\mathcal{T}_v$ to denote the sub parity decision tree starting with node v.
- The *depth* of a node is the number of its ancestors (e.g., the root has depth 0) and the depth of $\mathcal{T}$ is the maximum depth over all its leaves.
- We say that two parity decision trees $\mathcal{T}$ and $\mathcal{T}'$ are *equivalent* (denoted by $\mathcal{T} \equiv \mathcal{T}'$) if they compute the same function.

2.3 Noisy Decision Trees

▶ **Definition 12** (Noisy oracle). *A noisy query to a bit $b \in \{\pm 1\}$ with correlation $\gamma \in [-1,1]$ returns a bit $b' \in \{\pm 1\}$ where*

$$b' = \begin{cases} b & \textit{with probability } (1+\gamma)/2, \\ -b & \textit{with probability } (1-\gamma)/2. \end{cases}$$

The cost of a noisy query with correlation γ is defined to be γ^2.

▶ **Definition 13** (Noisy decision tree). *A* noisy decision tree $\mathcal{T}$ *is a rooted binary tree in which each internal node v is labeled by an index $q_v \in [n]$ and a correlation $\gamma_v \in [-1, 1]$. The outgoing edges are labeled by $+1$ and -1 and the leaves are labeled by 0 and 1.*

On input $x \in \{\pm 1\}^n$, the tree $\mathcal{T}$ constructs a computation path *$\mathcal{P}$ from the root to leaf as follows. When $\mathcal{P}$ reaches an internal node v, it makes a noisy query to x_{q_v} with correlation γ_v and follows the edge labeled by the outcome of this noisy query. The output of the tree is defined by sampling a root-to-leaf path and returning the label of the leaf. Since the computation path $\mathcal{P}$ is probabilistic, this is an inherently randomized model of computation. We use $\mathcal{T}(x) \in \{0, 1\}$ to denote the (probabilistic) output of $\mathcal{T}$ on input x. We also use $\mathcal{T}(v) \in \{0, 1\}$ to denote the label on v when v is a leaf. We do* not *require that the indices q_v queried along a path $\mathcal{P}$ are distinct. The* cost *of any path is the sum of costs of the noisy queries along that path; and the cost of $\mathcal{T}$ is the maximum cost of any root-to-leaf path.*

We remark that for any noisy decision tree $\mathcal{T}$, its Fourier coefficient $\widehat{\mathcal{T}}(S)$ is given by $\mathbb{E}[\mathcal{T}(x)x_S]$ where the expectation is over the randomness of both $x \sim \mathcal{U}_n$ and $\mathcal{T}$.

3 Useful Concentration Inequalities

We describe useful concentration inequalities in this section.

3.1 Low Degree Polynomials

We use the fact that low degree polynomials satisfy strong concentration properties under k-wise independent distributions. We will find the following hypercontractive inequality useful.

▶ **Theorem 14** ([5], see also [29, $(2, q)$-hypercontractivity]). *Let $f : \{\pm 1\}^n \to \mathbb{R}$ be a degree-d polynomial. Then for any $q \geq 2$, we have $\|f\|_q \leq (q-1)^{d/2} \|f\|_2$.*

▶ **Lemma 15.** *Let $f : \{\pm 1\}^n \to \mathbb{R}$ be a degree-d polynomial. Let $\mathcal{D}$ be a $2k$-wise independent distribution over $\{\pm 1\}^n$, where $k \geq d$. Let $\mu = \mathbb{E}_{x\sim\mathcal{D}}[f(x)]$ and $\sigma^2 = \mathbb{E}_{x\sim\mathcal{D}}\left[(f(x) - \mu)^2\right]$. Then for any $\alpha > 0$ and any integer $1 \leq \ell \leq k/d$, we have*

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[(f(x) - \mu)^{2\ell}\right] \leq \sigma^{2\ell} \cdot (2\ell - 1)^{d\cdot\ell}.$$

In particular we have

$$\mathop{\mathbf{Pr}}_{x\sim\mathcal{D}}\left[|f(x) - \mu| \geq \alpha \cdot \sigma\right] \leq \alpha^2 \cdot \left(\frac{2k}{d \cdot \alpha^{2/d}}\right)^k.$$

Proof. Observe that $(f(x) - \mu)^{2\ell}$ is a polynomial of degree at most $2\ell \cdot d \leq 2k$. Thus its expectation under $\mathcal{D}$ is the same as its expectation under the uniform distribution over $\{\pm 1\}^n$. By Theorem 14, we have

$$\|f - \mu\|_{2\ell} \leq (2\ell - 1)^{d/2} \|f - \mu\|_2 = \sigma \cdot (2\ell - 1)^{d/2}.$$

Hence by Markov's inequality, we have

$$\mathop{\mathbf{Pr}}_{x\sim\mathcal{D}}\left[|f(x) - \mu| \geq \alpha \cdot \sigma\right] \leq \frac{\mathbb{E}_{x\sim\mathcal{D}}\left[(f(x) - \mu)^{2\ell}\right]}{(\alpha \cdot \sigma)^{2\ell}} = \frac{\|f - \mu\|_{2\ell}^{2\ell}}{(\alpha \cdot \sigma)^{2\ell}} \leq \frac{(2\ell - 1)^{\ell\cdot d}}{\alpha^{2\ell}}.$$

Now we derive the second bound. We only need to focus on the case $\alpha \geq 1$ since otherwise the RHS is at least 1. Then by setting $\ell = \lfloor k/d \rfloor$, we have

$$\mathop{\mathbf{Pr}}_{x\sim\mathcal{D}}\left[|f(x) - \mu| \geq \alpha \cdot \sigma\right] \leq \frac{(2\lfloor k/d \rfloor - 1)^{\lfloor k/d \rfloor \cdot d}}{\alpha^{2\lfloor k/d \rfloor}} \leq \frac{(2k/d)^k}{\alpha^{2(k/d - 1)}} = \alpha^2 \cdot \left(\frac{2k}{d \cdot \alpha^{2/d}}\right)^k.$$ ◀

3.2 Martingales

We show an adaptive version of Azuma's inequality for martingales. The proof is similar to the inductive proof of the standard Azuma's inequality and thus deferred to Appendix B.

▶ **Lemma 16** (Adaptive Azuma's inequality). *Let $X^{(0)},\dots,X^{(D)}$ be a martingale and $\Delta^{(1)},\dots,\Delta^{(D)}$ be a sequence of magnitudes such that $X^{(0)}=0$ and $X^{(i)}=X^{(i-1)}+\Delta^{(i)}\cdot z^{(i)}$ for $i\in[D]$, where if conditioning on $z^{(1)},\dots,z^{(i-1)}$,*
(1) *$z^{(i)}$ is a mean-zero random variable and $\left|z^{(i)}\right|\le 1$ always holds;*
(2) *$\Delta^{(i)}$ is a fixed value.*
If there exists some constant $U\ge 0$ such that $\sum_{i=1}^{D}\left|\Delta^{(i)}\right|^2\le U$ always holds, then for any $\beta\ge 0$ we have

$$\mathbf{Pr}\left[\max_{i=0,1,\dots,D}\left|X^{(i)}\right|\ge\beta\cdot\sqrt{2U}\right]\le 2\cdot e^{-\beta^2/2}.$$

Next, we generalize Lemma 16 as follows.

▶ **Lemma 17.** *Let $m\ge 1$ be an integer. For each $t\in[m]$, let $X_t^{(0)},\dots,X_t^{(D)}$ be a sequence of random variables and $\Delta_t^{(1)},\dots,\Delta_t^{(D)}$ be a sequence of magnitudes such that $X_t^{(0)}=0$ and $X_t^{(i)}=X_t^{(i-1)}+\Delta_t^{(i)}\cdot z_t^{(i)}+\mu_t^{(i)}$ for $i\in[D]$, where if conditioning on $z_t^{(1)},\dots,z_t^{(i-1)}$,*
(1) *$z_t^{(i)}$ is a mean-zero random variable and $\left|z_t^{(i)}\right|\le 1$ always holds;*
(2) *$\Delta_t^{(i)}$ is a fixed value and $\mu_t^{(i)}$ is a random variable.*
If there exist some constants $U,V\ge 0$ and $\eta\in[0,1]$ such that

$$\mathbf{Pr}\left[\exists t\in[m],\ \left(\sum_{i=1}^{D}\left|\Delta_t^{(i)}\right|^2>U\right)\vee\left(\sum_{i=1}^{D}\left|\mu_t^{(i)}\right|>V\right)\right]\le\eta,$$

then for any $\beta\ge 0$ we have

$$\mathbf{Pr}\left[\exists t\in[m],\ \max_{i=0,1,\dots,D}\left|X_t^{(i)}\right|\ge V+\beta\cdot\sqrt{2U}\right]\le\eta+2m\cdot e^{-\beta^2/2}.$$

Proof. We divide the proof into the following two cases.

Case $\eta=0$. Let $\widehat{X}_t^{(i)}=X_t^{(i)}-\sum_{j=1}^{i}\mu_t^{(j)}$ for each t and i. Then $\left|X_t^{(i)}\right|=\left|\widehat{X}_t^{(i)}+\sum_{j=1}^{i}\mu_t^{(j)}\right|\le V+\left|\widehat{X}_t^{(i)}\right|$. By a union bound, it suffices to show for any fixed t, we have

$$\mathbf{Pr}\left[\max_{i=0,1,\dots,D}\left|\widehat{X}_t^{(i)}\right|\ge\beta\cdot\sqrt{2U}\right]\le 2\cdot e^{-\beta^2/2},$$

which follows from Lemma 16.

Case $\eta\ge 0$. Consider $\widetilde{X}_t^{(0)},\dots,\widetilde{X}_t^{(D)}$ defined by setting $\widetilde{X}_t^{(0)}=0$ and $\widetilde{X}_t^{(i)}=\widetilde{X}_t^{(i-1)}+\widetilde{\Delta}_t^{(i)}\cdot z_t^{(i)}+\widetilde{\mu}_t^{(i)}$, where

$$\widetilde{\Delta}_t^{(i)}=\begin{cases}\Delta_t^{(i)} & \sum_{j=1}^{i}\left|\Delta_t^{(j)}\right|^2\le U,\\ 0 & \text{otherwise,}\end{cases}\quad\text{and}\quad \widetilde{\mu}_t^{(i)}=\begin{cases}\mu_t^{(i)} & \sum_{j=1}^{i}\left|\mu_t^{(j)}\right|\le V,\\ 0 & \text{otherwise.}\end{cases}$$

Then Item (1) and (2) hold for $\left(\widetilde{X}_t^{(i)}\right)_{t,i}$ and $\left(\widetilde{\Delta}_t^{(i)}\right)_{t,i},\left(\widetilde{\mu}_t^{(i)}\right)_{t,i}$.

Note that $\mathbf{Pr}\left[\exists t \in [m], i \in \{0,1\ldots,D\}, \widetilde{X}_t^{(i)} \neq X_t^{(i)}\right] \le \eta$ and $\sum_{i=1}^{D}\left|\widetilde{\Delta}_t^{(i)}\right|^2 \le U$, $\sum_{i=1}^{D}\left|\widetilde{\mu}_t^{(i)}\right| \le V$ always. Hence from the previous case, we have

$$\begin{aligned}
&\mathbf{Pr}\left[\exists t \in [m], \max_{i=0,1,\ldots,D}\left|X_t^{(i)}\right| \ge V + \beta\cdot\sqrt{2U}\right] \\
&\le \mathbf{Pr}\left[\exists t \in [m], i \in \{0,1\ldots,D\},\ \widetilde{X}_t^{(i)} \neq X_t^{(i)}\right] \\
&\quad + \mathbf{Pr}\left[\exists t \in [m], \max_{i=0,1,\ldots,D}\left|\widetilde{X}_t^{(i)}\right| \ge V + \beta\cdot\sqrt{2U}\right] \\
&\le \eta + 2m\cdot e^{-\beta^2/2}.
\end{aligned}$$

◀

4 How to Clean Up Parity Decision Trees

In this section we show how to *clean up* the given parity decision tree to make it easier to analyze.

4.1 k-cleanness

It will be useful to identify $\mathbb{F}_2^n$ with $\{\pm1\}^n$ by $\mathsf{Enc}\colon (x_1,\ldots,x_n) \mapsto ((-1)^{x_1},\ldots,(-1)^{x_n})$. For a subset $X \subseteq \mathbb{F}_2^n$ we will denote $\mathsf{Enc}(X) = \{\mathsf{Enc}(x) : x \in X\}$. Thus, we may think of Boolean functions also as $f\colon \mathbb{F}_2^n \to \{0,1\}$. We observe that under this representation of the input, a parity decision tree $\mathcal{T} : \mathbb{F}_2^n \to \{0,1\}$ indeed queries parity functions (i.e., linear functions over $\mathbb{F}_2$) of the input bits $x \in \mathbb{F}_2^n$ and decides whether to go left or right based on their outcome. Thus, the set of all possible inputs in $\mathbb{F}_2^n$ that reach a given node in a parity decision tree is an affine subspace of $\mathbb{F}_2^n$.

We introduce some notation.

▶ **Notation 18.** Let $\mathcal{T}\colon \{\pm1\}^n \to \{0,1\}$ be a parity decision tree and let v be a node in it.

- We use $\mathcal{P}_v \subseteq \{\pm1\}^n$ to denote the set of all points reaching node v. Note that $\mathcal{P}_v = \mathsf{Enc}(H_v + a)$ where H_v is a linear subspace of $\mathbb{F}_2^n$ of dimension $n - \mathsf{depth}(v)$ and $a \in \mathbb{F}_2^n$.
- For any $S \subseteq [n]$, we define $\widehat{\mathcal{P}_v}(S) = \mathbb{E}_{x\sim\mathcal{P}_v}[x_S]$.
- We use $\mathcal{S}_v$ to denote all fully correlated sets with $\mathcal{P}_v$, i.e., $\mathcal{S}_v = \left\{S \subseteq [n] \,\middle|\, \widehat{\mathcal{P}_v}(S) \in \{\pm1\}\right\}$. We observe that if $\mathcal{P}_v = \mathsf{Enc}(H_v + a)$, then $\mathcal{S}_v = H_v^{\perp}$. Additionally, if the queries on the path from root to v are $Q_{v_0},\ldots,Q_{v_{i-1}}$, then $\mathcal{S}_v = \mathsf{Span}\langle\{Q_{v_0},\ldots,Q_{v_{i-1}}\}\rangle$.
- If v is an internal node, then define $J(v)$ as the set of newly fixed coordinates after querying Q_v, i.e., $i \in J(v)$ iff $\{i\} \notin \mathcal{S}_v$ but $\{i\} \in \mathsf{Span}\langle\mathcal{S}_v \cup \{Q_v\}\rangle$.

The following simple fact shows that there is no "somewhat" correlated set.

▶ **Fact 19.** *For any parity decision tree $\mathcal{T}$ and any node v in $\mathcal{T}$, $\widehat{\mathcal{P}_v}(S) \in \{+1,0,-1\}$ holds for any set S.*

Proof. Since $\mathcal{P}_v = \mathsf{Enc}(H_v + a)$ where $H_v + a$ is an affine subspace, $\mathcal{P}_v$ falls into one of the following 3 cases: (a) all points in $\mathcal{P}_v$ satisfy $\chi_S(x) = 1$, (b) all points satisfy $\chi_S(x) = -1$, (c) exactly half of the points satisfy $\chi_S(x) = 1$. ◀

Let $\mathcal{S} \subseteq \mathbb{F}_2^n$ be a subspace and $S \subseteq [n]$. For simplicity, we write $S \in \mathcal{S}$ iff the indicator vector of S is contained in $\mathcal{S}$. Now we describe the desired property: *k-clean*.

▶ **Definition 20** (k-clean subspace and mess-witness). *Let k be a positive integer. A subspace $\mathcal{S}$ is k-*clean *if for any set $S \in \mathcal{S}$ such that $|S| \leq k$, we have that $\{i\} \in \mathcal{S}$ holds for any $i \in S$.*

Moreover, when $\mathcal{S}$ is not k-clean, we say i is a mess-witness *if there exists some $S \ni i, |S| \leq k$ such that $S \in \mathcal{S}$ but $\{i\} \notin \mathcal{S}$.*

▶ **Definition 21** (k-clean parity decision tree). *A parity decision tree $\mathcal{T}$ is k-*clean *if the following holds:*

- *For any internal node v, either (a) $\mathcal{S}_v$ is k-clean, or (b) $Q_v = \{i\}$ where i is a mess-witness for $\mathcal{S}_v$. Moreover, we say v is k-*clean *if (a) holds; and we say v is* cleaning *if (b) holds.*
- *For any leaf v, $\mathcal{S}_v$ is k-clean (in such a case, we say that v is k-*clean*).*
- *For any k-clean internal node v, $\mathcal{T}_v$ starts with $\ell(v)$ non-adaptive queries[6] where $\ell(v) \geq 1$. In addition, for any $i \in \{1, \ldots, \ell(v) - 1\}$, any node of depth i in $\mathcal{T}_v$ is cleaning; and all node of depth $\ell(v)$ are k-clean.[7]*

▶ **Example 22.** If $\mathcal{T}$ is a decision tree (i.e., $|Q_v| \equiv 1$ for any internal node v) then it is k-clean for any k, where each internal node is k-clean.

If $\mathcal{T}$ is the depth-1 parity decision tree for $\mathcal{T}(x) = x_1x_2x_3$ (i.e., $\mathcal{T}$ only has a root v_0 querying $Q_{v_0} = \{1, 2, 3\}$), then it is 2-clean but not 3-clean, since for either leaf v we have $\{1, 2, 3\} \in \mathcal{S}_v$ but $\{1\} \notin \mathcal{S}_v$.

The benefit of having a k-clean parity decision tree is that it makes the expression of Fourier coefficients simpler.

▶ **Lemma 23.** *Let $\mathcal{T} \colon \{\pm 1\}^n \to \{0, 1\}$ be a k-clean parity decision tree and let S be a set of size $\ell \leq k$. Let $v_0, \ldots, v_d$ be a random root-to-leaf path. Define $\boldsymbol{v}^{(0)}, \ldots, \boldsymbol{v}^{(d)} \in \{-1, 0, +1\}^n$ by setting $\boldsymbol{v}_j^{(i)} = \widehat{\mathcal{P}_{v_i}}(j)$ for each i, j. Recall that $\boldsymbol{v}_S^{(d)} = \prod_{j \in S} \boldsymbol{v}_j^{(d)}$. Then we have*

$$\widehat{\mathcal{T}}(S) = \mathop{\mathbb{E}}_{v_0, \ldots, v_d}\left[\mathcal{T}(v_d) \cdot \boldsymbol{v}_S^{(d)}\right].$$

Proof. Observe that for any $j \in J(v_i) \subseteq J$, the j-th coordinate is fixed after querying Q_{v_i}. Therefore we have

$$\widehat{\mathcal{T}}(S) = \mathop{\mathbb{E}}_{y \sim \mathcal{U}_n}[\mathcal{T}(y) \cdot y_S] = \mathop{\mathbb{E}}_{v_0, \ldots, v_d}\left[\mathcal{T}(v_d) \cdot \mathop{\mathbb{E}}_{y \sim \mathcal{P}_{v_d}}[y_S]\right] = \mathop{\mathbb{E}}_{v_0, \ldots, v_d}\left[\mathcal{T}(v_d) \cdot \widehat{\mathcal{P}}_{v_d}(S)\right]$$

By Fact 19, $\widehat{\mathcal{P}}_{v_d}(S) \neq 0$ iff $S \in \mathcal{S}_{v_d}$, which, due to $\ell \leq k$ and v_d being a k-clean leaf, is equivalent to all coordinates in S being fixed along this path. Hence $\widehat{\mathcal{P}}_{v_d}(S) = \prod_{j \in S} \boldsymbol{v}_j^{(d)}$. ◀

4.2 Cleanup Process

We first analyze the cleanup process for a subspace.[8]

---

[6] This means for any $i \in \{0, 1 \ldots, \ell(v) - 1\}$, all nodes of depth i in $\mathcal{T}_v$ make the same query.

[7] This "leveled adaptive" condition is required just for convenience of proofs. In fact, one can show that the first few queries in $\mathcal{T}_v$ can be rearranged to make sure they are non-adaptive until we reach a k-clean node. See Lemma 24.

[8] The $k = 2$ case of Lemma 24 is essentially [4, Proposition 3.5]. However there is a gap in their proof. For example, if the parity decision tree non-adaptively queries $x_1x_2x_3x_4, x_1x_5, x_2x_6$ in order, then their analysis fails.

▶ **Lemma 24** (Clean subspace). *Let $k \geq 2$ be an integer and $\mathcal{S}$ be a subspace of rank at most d. We construct a new subspace $\mathcal{S}'$ (initialized as $\mathcal{S}$) as follows: while $\mathcal{S}'$ is not k-clean, we continue to update $\mathcal{S}' \leftarrow \mathsf{Span}\langle \mathcal{S}' \cup \{\{i\}\}\rangle$ with some mess-witness i. Then $\mathsf{rank}(\mathcal{S}') \leq d \cdot k$ and any update choice of mess-witnesses will result in the same final subspace $\mathcal{S}'$.*

Proof. Assume $\mathcal{S}$ is a subspace of $\mathbb{F}_2^n$. Then first note that the number of updates is finite, since we can update for at most n times.

Next we show that the number of updates and the final $\mathcal{S}'$ does not depend on the choice of mess-witnesses. We do so by an exchange argument. Let $i_1, \ldots, i_r$ and $i'_1, \ldots, i'_{r'}$ be two rounds of execution using different mess-witnesses. Then there exists some $t < \min\{r, r'\}$ such that $i_j = i'_j$ for all $j \leq t$, but $i_{t+1} \neq i'_{t+1}$. Let $\mathcal{S}_t = \mathsf{Span}\langle \mathcal{S} \cup \{\{i_1\}, \ldots, \{i_t\}\}\rangle$. Then there exist $S \ni i_{t+1}$ and $S' \ni i'_{t+1}$ (possibly $S = S'$) such that $S, S' \in \mathcal{S}_t$ but $\{i_{t+1}\}, \{i'_{t+1}\} \notin \mathcal{S}_t$. Since the final subspace is k-clean, we know there exists some $T \geq t$ such that

$$\{i_{t+1}\} \notin \mathsf{Span}\langle \mathcal{S} \cup \{\{i'_1\}, \ldots, \{i'_T\}\}\rangle \quad \text{but} \quad \{i_{t+1}\} \in \mathsf{Span}\left\langle \mathcal{S} \cup \left\{\{i'_1\}, \ldots, \{i'_{T+1}\}\right\}\right\rangle,$$

which means $\left\{i'_{T+1}, i_{t+1}\right\} \in \mathsf{Span}\langle \mathcal{S} \cup \{\{i'_1\}, \ldots, \{i'_T\}\}\rangle$. Hence we can safely replace i'_{T+1} with i_{t+1}, and then swap i_{t+1} with i'_{t+1}. We can perform this process as long as $(i_1, \ldots, i_r) \neq (i'_1, \ldots, i'_{r'})$, which means $r = r'$ and the final $\mathcal{S}'$ is always the same.

For any subspace $\mathcal{H}$, we define $\mathsf{rank}_1(\mathcal{H}) = |\{i \mid \{i\} \in \mathcal{H}\}|$ and thus $\mathsf{rank}(\mathcal{H}) - \mathsf{rank}_1(\mathcal{H}) \geq 0$. Now we analyze the following particular way to construct $\mathcal{S}'$: We initialize $\mathcal{S}'$ as $\mathcal{S}$. While $\mathcal{S}'$ is not k-clean, we find a minimal $S = \{i_1, \ldots, i_s\} \in \mathcal{S}'$ such that i_1 is a mess-witness; then we update $\mathcal{S}' \leftarrow \mathsf{Span}\langle \mathcal{S}' \cup \{\{i_1\}, \ldots, \{i_{s-1}\}\}\rangle$. Note that before the update, $1 < s \leq k$ and $\{i_j\} \notin \mathcal{S}'$ holds for each $j \in [s]$, since S is minimal and $\mathcal{S}'$ is not k-clean. Thus after the update, $\mathsf{rank}(\mathcal{S}')$ grows by $s - 1 \leq k - 1$ and $\mathsf{rank}_1(\mathcal{S}')$ grows by s, which means $\mathsf{rank}(\mathcal{S}') - \mathsf{rank}_1(\mathcal{S}')$ shrinks by 1. Hence we have at most $\mathsf{rank}(\mathcal{S}) - \mathsf{rank}_1(\mathcal{S}) \leq d$ updates before $\mathcal{S}'$ is k-clean; and the final $\mathcal{S}'$ has rank at most $\mathsf{rank}(\mathcal{S}) + (k-1) \cdot d \leq d \cdot k$. ◀

We now show how to convert an arbitrary parity decision tree into a k-clean parity decision tree which still has a small depth and fixes a small number of variables along each path. The latter quantity is in fact bounded by the depth as shown in Fact 25.

▶ **Fact 25.** *Let $\mathcal{T}$ be a depth-d parity decision tree. Let $v_0, \ldots, v_{d'}$ be any root-to-leaf path. Then we have $\sum_{i=0}^{d'-1} |J(v_i)| \leq d'$.*

Proof. Observe that $\sum_{i=0}^{d'-1} |J(v_i)| = \left|\left\{i \,\middle|\, \{i\} \in \mathsf{Span}\left\langle Q_{v_0}, \ldots, Q_{v_{d'-1}}\right\rangle\right\}\right| \leq d'$. ◀

▶ **Corollary 26.** *Let $\mathcal{T}$ be a depth-D k-clean parity decision tree. Let $v_0, \ldots, v_{D'}$ be any root-to-leaf path where at most d of the nodes $v_0, \ldots, v_{D'-1}$ are k-clean. Then $\sum_{i:|J(v_{i-1})|>1} |J(v_i)| \leq 2d$.*

Proof. By Fact 25 we have $\sum_{i=0}^{D'-1} |J(v_i)| - 1 \leq 0$. Since any v_i with $J(v_i) = \emptyset$ is not cleaning and therefore must be k-clean. Thus

$$\sum_{i:|J(v_i)|>1} |J(v_i)| - 1 \leq |\{i : J(v_i) = \emptyset\}| \leq d.$$

For $|J(v_i)| > 1$, we have $|J(v_i)| - 1 \geq |J(v_i)|/2$ and thus $\sum_{i:|J(v_i)|>1} |J(v_i)| \leq 2d$. ◀

▶ **Lemma 27** (Clean parity decision tree). *Let $k \geq 2$ be an integer. Let $\mathcal{T}$ be an arbitrary depth-d parity decision tree. Then there exists a k-clean parity decision tree $\mathcal{T}'$ of depth at most $d \cdot k$ equivalent to $\mathcal{T}$. Moreover, any root-to-leaf path in $\mathcal{T}'$ has at most d nodes that are k-clean.*

Algorithm 1 Clean parity decision tree: build $\mathcal{T}'$ from $\mathcal{T}$.

Input: an arbitrary depth-d parity decision tree $\mathcal{T}$
Output: a parity decision tree $\mathcal{T}'$ with desired properties
1 $r \leftarrow$ root of $\mathcal{T}$
2 Initialize the root of $\mathcal{T}'$ as r'
3 `Build`$(r, r', 1)$
4 **Procedure** `Build`(v, v', ℓ)
  /* (v, v') are the current nodes on $(\mathcal{T}, \mathcal{T}')$; ℓ is the recursion depth. */
5   **if** v *is a leaf* **then** Label v' with the label of v
6   **else**
7     $(v_-, v_+) \leftarrow$ the left and right child of v
8   **if** $\widehat{\mathcal{P}_{v'}}(Q_v) = -1$ **then** `Build`$(v_-, v', \ell + 1)$
9   **else if** $\widehat{\mathcal{P}_{v'}}(Q_v) = +1$ **then** `Build`$(v_+, v', \ell + 1)$
10   **else** /* $\widehat{\mathcal{P}_{v'}}(Q_v) = 0$ due to Fact 19 */
11       $Q_{v'} \leftarrow Q_v$
12       $(v'_-, v'_+) \leftarrow$ the left and right child of v'
13   Initialize $O \leftarrow \emptyset$
14   **while** $\mathsf{Span}\left\langle \mathcal{S}_{v'} \cup \{Q_{v'}\} \cup O \right\rangle$ *is not k-clean* **do**
15   Update $O \leftarrow O \cup \{\{i\}\}$, where i is a *mess-witness*
16   **end**
17       $\mathcal{T}'$ non-adaptively queries every set (which is a singleton) in O under v' in arbitrary order
18   **foreach** *leaf* $\widehat{v}$ *under* v'_- **do** `Build`$(v_-, \widehat{v}, \ell + 1)$
19   **foreach** *leaf* $\widehat{v}$ *under* v'_+ **do** `Build`$(v_+, \widehat{v}, \ell + 1)$
20
21   **end**
22   **end**

Proof. We build $\mathcal{T}'$ by the following recursive algorithm. An example of the algorithm is provided in Figure 1

We now prove the correctness of Algorithm 1, which is guaranteed by the following claims.

- **For any internal node $v' \in \mathcal{T}'$, $Q_{v'}$ is not implied by its ancestors' queries.** By Fact 19, this is equivalent to $Q_{v'} \notin \mathcal{S}_{v'}$, which follows from the conditions in Line 8/9/13.
- **The depth of $\mathcal{T}'$ is at most $d \cdot k$.** Let $v_0, \ldots, v_{d'}$ be any root-to-leaf path of $\mathcal{T}$ and let $\mathcal{P}'$ be its corresponding path in $\mathcal{T}'$. Then the construction process of $\mathcal{P}'$ corresponds to the cleanup process for $\mathsf{Span}\left\langle Q_{v_0}, \ldots, Q_{v_{d'-1}} \right\rangle$ in Lemma 24; hence the depth of $\mathcal{T}'$ equals $\mathsf{rank}(\mathcal{S}') \le d' \cdot k \le d \cdot k$ where $\mathcal{S}'$ is the k-clean subspace produced by applying Lemma 24.
- **$\mathcal{T} \equiv \mathcal{T}'$ and any root-to-leaf path in $\mathcal{T}'$ has at most d k-clean nodes.** This is because $\mathcal{T}'$ only refines $\mathcal{T}$ by inserting cleaning nodes.
- **Whenever we call `Build`$(\cdot, v', \cdot)$, v' is k-clean.** We prove by induction on ℓ. The base case Line 3 is obvious. For Line 8/9, we recurse on the same v', which is k-clean by induction. For Line 17/18, note that $\mathcal{S}_{\widehat{v}} = \mathsf{Span}\left\langle \mathcal{S}_{v'} \cup \{Q_{v'}\} \cup O \right\rangle$; hence from the condition in Line 13, it is k-clean.
- **Nodes created in Line 16 are cleaning.** Let $o = |O|$ and let $i_1, i_2, \ldots, i_o$ be the query

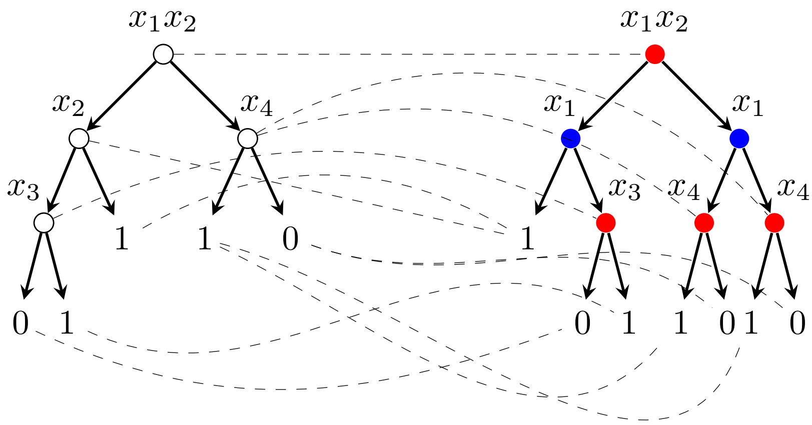

Figure 1 An example of the cleanup process with $k = 2$ where the LHS is $\mathcal{T}$ and the RHS is $\mathcal{T}'$. All the left (resp., right) outgoing edges are labeled with -1 (resp., $+1$). Red nodes and leaves are k-clean, and blue nodes are cleaning (i.e., non-adaptive queries). Nodes connected with dashed curves are invoked by Build.

order. For any $j \in [o]$, let v'_j be any one of the nodes created for i_j, then

$$\mathcal{S}_{v'_j} = \mathsf{Span}\left\langle \mathcal{S}_{v'} \cup \{Q_{v'}\} \cup \{\{i_1\}, \ldots, \{i_{j-1}\}\}\right\rangle,$$

which is not k-clean by Line 13; hence v'_j is cleaning by the condition in Line 13. ◀

5 Fourier Bounds for Parity Decision Trees

Our goal in this section is to prove Theorem 1 with detailed bounds provided.

5.1 Level-1 Bound

We first prove the concentration result for level-1. We start with the following simple bound for general parity decision trees.

▶ **Lemma 28.** *Let $\mathcal{T}\colon \{\pm 1\}^n \to \{0,1\}$ be a depth-D parity decision tree. Let $v_0, \ldots, v_{D'}$ be any root-to-leaf path. Define $\boldsymbol{v}^{(0)}, \ldots, \boldsymbol{v}^{(D')} \in \{-1, 0, +1\}^n$ by setting $\boldsymbol{v}_j^{(i)} = \widehat{\mathcal{P}_{v_i}}(j)$ for each $0 \le i \le D'$ and $j \in [n]$. Then for any $a_1, \ldots, a_n \in \{-1, 0, 1\}$, we have $\left|\sum_{j=1}^n a_j \cdot \boldsymbol{v}_j^{(D')}\right| \le D' \le D$.*

Proof. Note that the set of non-zero coordinates in $\boldsymbol{v}^{(D')}$ is exactly $\bigcup_{i=0}^{D'-1} J(v_i)$. Hence by Fact 25, we have

$$\left|\sum_{j=1}^{n} a_j \cdot \boldsymbol{v}_j^{(D')}\right| \le \sum_{j=1}^{n} \left|\boldsymbol{v}_j^{(D')}\right| = \sum_{i=0}^{D'-1} |J(v_i)| \le D' \le D. \qquad ◀$$

Now we give an improved bound for k-clean parity decision trees. To do so, we need one more notation which will be crucial in our analysis.

▶ **Notation 29.** Let $\mathcal{T}$ be a k-clean parity decision tree. For any node v, we define $C(v)$ as the nearest ancestor of v (including itself) that is k-clean.

▶ **Lemma 30.** *There exists a universal constant $\kappa \geq 1$ such that the following holds. Let $\mathcal{T}\colon \{\pm 1\}^n \to \{0,1\}$ be a depth-D $2k$-clean parity decision tree where $k \geq 1$ and any root-to-leaf path has at most d nodes that are $2k$-clean.*

Let $v_0, \ldots, v_{D'}$ be a random root-to-leaf path. Define $\boldsymbol{v}^{(0)}, \ldots, \boldsymbol{v}^{(D')} \in \{-1,0,+1\}^n$ by setting $\boldsymbol{v}^{(i)}_j = \widehat{\mathcal{P}_{v_i}}(j)$ for each $0 \leq i \leq D'$ and $j \in [n]$. Then for any $a_1, \ldots, a_n \in \{-1,0,1\}$ and any $\varepsilon \leq 1/2$, we have $\mathbf{Pr}\left[\left|\sum_{j=1}^n a_j \cdot \boldsymbol{v}^{(D')}_j\right| \geq R(D,d,k,\varepsilon)\right] \leq \varepsilon$, where

$$R(D,d,k,\varepsilon) = \kappa \cdot \sqrt{\left(D + dk\left(\frac{1}{\varepsilon}\right)^{\frac{1}{k}}\right)\log\left(\frac{1}{\varepsilon}\right)}.$$

In the proof of Lemma 30 we will use the following simple claim.

▶ **Fact 31.** *Let $p_1, \ldots, p_n$ be a sub-probability distribution, i.e., $p_i \geq 0$ and $\sum_{i=1}^n p_i \leq 1$. Let $a_1, \ldots, a_n \in \mathbb{R}$. Then for any $k \in \mathbb{N}$, we have $\sum_{i=1}^n p_i a_i^{2k} \geq \left(\sum_{i=1}^n p_i a_i^2\right)^k$.*

Proof. We add $p_{n+1} = 1 - (\sum_{i=1}^n p_i)$ and $a_{n+1} = 0$ so p is a probability distribution. Then the claim follows from $\mathbb{E}[X^k] \geq \mathbb{E}[X]^k$, where random variable X gets value a_i^2 with probability p_i. ◀

Proof of Lemma 30. Extend $\boldsymbol{v}^{(D'+1)} = \cdots = \boldsymbol{v}^{(D)}$ to equal $\boldsymbol{v}^{(D')}$. For each $0 \leq i \leq D$, let $X^{(i)} = \sum_{j=1}^n a_j \cdot \boldsymbol{v}^{(i)}_j$. We define $\delta^{(i)} = 0$ for $D' < i \leq D$. For $1 \leq i \leq D'$, we let

$$\delta^{(i)} = X^{(i)} - X^{(i-1)} = \sum_{j=1}^n a_j \cdot \left(\boldsymbol{v}^{(i)}_j - \boldsymbol{v}^{(i-1)}_j\right) = \sum_{j \in J(v_{i-1})} a_j \cdot \boldsymbol{v}^{(i)}_j,$$

where $J(v_{i-1})$ depends only on $C(v_{i-1})$ since $\mathcal{T}_{C(v_{i-1})}$ performs non-adaptive queries before (and possibly even after) reaching v_i. Note that for the two possible outcomes of querying Q_{v_i}, $\boldsymbol{v}^{(i)}_j$ is fixed to ± 1 respectively for each $j \in J(v_{i-1})$. Thus $\delta^{(i)} = \Delta^{(i)} \cdot z^{(i)}$ where $\Delta^{(i)}$ is a fixed value given $z^{(1)}, \ldots, z^{(i-1)}$ and $z^{(1)}, \ldots, z^{(D')}$ are independent unbiased coins in $\{\pm 1\}$.

Since $C(v_{i-1})$ is $2k$-clean, the collection of random variables $\left\{\boldsymbol{v}^{(i)}_j \,\middle|\, j \in J(v_{i-1})\right\}$ is $2k$-wise independent conditioning on $C(v_{i-1})$. Note that δ_i is a linear function and

$$\mathbb{E}\left[\delta^{(i)} \,\middle|\, C(v_{i-1})\right] = 0 \quad \text{and} \quad \mathbb{E}\left[\left(\delta^{(i)}\right)^2 \,\middle|\, C(v_{i-1})\right] = \sum_{j \in J(v_{i-1})} a_j^2 \leq |J(v_{i-1})|\,.$$

By the first bound in Lemma 15, we have

$$\mathbb{E}\left[\left(\delta^{(i)}\right)^{2k} \,\middle|\, C(v_{i-1})\right] \leq (2k-1)^k \cdot |J(v_{i-1})|^k\,, \tag{5}$$

and $\left|\delta^{(i)}\right| \leq |J(v_{i-1})|$ always. Our first goal is to bound $\mathbf{Pr}\left[\sum_{i=1}^D \left(\delta^{(i)}\right)^2 > D + 2\alpha^2 d\right]$. Observe that whenever the event $\sum_{i=1}^D \left(\delta^{(i)}\right)^2 > D + 2\alpha^2 d$ happens, it must be the case that $\sum_{i:|J(v_{i-1})|>1} \left(\delta^{(i)}\right)^2 > 2\alpha^2 d$. Thus,

$$
\begin{aligned}
\mathbf{Pr}\left[\sum_{i=1}^{D}\left(\delta^{(i)}\right)^2 > D + 2\alpha^2 d\right] &\le \mathbf{Pr}\left[\sum_{i:|J(v_{i-1})|>1}\left(\delta^{(i)}\right)^2 > 2\alpha^2 d\right] \\
&= \mathbf{Pr}\left[\sum_{i:|J(v_{i-1})|>1}\frac{|J(v_{i-1})|}{2d}\cdot\frac{\left(\delta^{(i)}\right)^2}{|J(v_{i-1})|} > \alpha^2\right] \\
&\le \mathbf{Pr}\left[\sum_{i:|J(v_{i-1})|>1}\frac{|J(v_{i-1})|}{2d}\cdot\frac{\left(\delta^{(i)}\right)^{2k}}{|J(v_{i-1})|^k} > \alpha^{2k}\right] \\
&\qquad\qquad\text{(by Fact 31 and Corollary 26)} \\
&= \mathbf{Pr}\left[\sum_{i:|J(v_{i-1})|>1}\frac{\left(\delta^{(i)}\right)^{2k}}{|J(v_{i-1})|^{k-1}} > 2d\cdot\alpha^{2k}\right] \\
&\le \mathbb{E}\left[\sum_{i:|J(v_{i-1})|>1}\frac{\left(\delta^{(i)}\right)^{2k}}{|J(v_{i-1})|^{k-1}}\right]\cdot\frac{1}{2d\cdot\alpha^{2k}}. \\
&\qquad\qquad\text{(by Markov's inequality)}
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
\mathbb{E}\left[\sum_{i:|J(v_{i-1})|>1}\frac{\left(\delta^{(i)}\right)^{2k}}{|J(v_{i-1})|^{k-1}}\right] &= \sum_{i=1}^{D}\mathop{\mathbb{E}}_{C(v_{i-1})}\left[\frac{1_{|J(v_{i-1})|>1}}{|J(v_{i-1})|^{k-1}}\cdot\mathbb{E}\left[\left(\delta^{(i)}\right)^{2k}\,\middle|\,C(v_{i-1})\right]\right] \\
&\le \sum_{i=1}^{D}\mathop{\mathbb{E}}_{C(v_{i-1})}\left[1_{|J(v_{i-1})|>1}\cdot(2k-1)^k\cdot|J(v_{i-1})|\right] \qquad \text{(by (5))} \\
&= (2k-1)^k\cdot\mathbb{E}\left[\sum_{i:|J(v_{i-1}|>1}|J(v_{i-1})|\right] \\
&\le (2k-1)^k\cdot 2d. \qquad\qquad \text{(by Corollary 26)}
\end{aligned}
$$

Overall, we have

$$
\mathbf{Pr}\left[\sum_{i=1}^{D}\left(\delta^{(i)}\right)^2 > D + 2\alpha^2 d\right] \le \frac{(2k-1)^k}{\alpha^{2k}}.
$$

Then by Lemma 17 with $m = 1$, we have

$$
\mathbf{Pr}\left[\left|X^{(D)}\right| = \left|\sum_{j=1}^{n} a_j\cdot \boldsymbol{v}_j^{(D)}\right| \ge \beta\sqrt{2\cdot(D+2\alpha^2 d)}\right] \le 2\cdot e^{-\beta^2/2} + \frac{(2k-1)^k}{\alpha^{2k}}.
$$

The desired bound follows from setting

$$
\alpha = \left(\frac{2}{\varepsilon}\right)^{\frac{1}{2k}}\sqrt{2k-1}, \quad \text{and} \quad \beta = \Theta\left(\sqrt{\log\left(\frac{1}{\varepsilon}\right)}\right). \qquad ◀
$$

Now we prove the complete level-1 bound for parity decision trees.

▶ **Theorem 32.** *Let $\mathcal{T}\colon \{\pm 1\}^n \to \{0,1\}$ be a depth-d parity decision tree. Let $p = \mathbf{Pr}[\mathcal{T}(x)=1] \in [2^{-d}, 1/2]$.*[9] *Then we have*

$$\sum_{j=1}^{n} \left|\widehat{\mathcal{T}}(j)\right| \le p \cdot \min\left\{d, O\left(\sqrt{d}\cdot \log\left(\frac{1}{p}\right)\right)\right\} = O\left(\sqrt{d}\right).$$

Proof. For any $i \in [n]$, let $a_i = \mathsf{sgn}\left(\widehat{\mathcal{T}}(i)\right)$. Now we prove the two bounds separately.

First Bound. Let $v_0, \ldots, v_{d'}$ be a random root-to-leaf path in $\mathcal{T}$. Define $\boldsymbol{v}^{(0)}, \ldots, \boldsymbol{v}^{(d')} \in \{-1,0,+1\}^n$ by setting $\boldsymbol{v}_j^{(i)} = \widehat{\mathcal{P}_{v_i}}(j)$ for each $0 \le i \le d'$ and $j \in [n]$. Since $\mathcal{T}$ is 1-clean in itself, by Lemma 23 we have

$$\sum_{j=1}^{n} \left|\widehat{\mathcal{T}}(j)\right| = \sum_{j=1}^{n} a_i \cdot \widehat{\mathcal{T}}(j) = \mathop{\mathbb{E}}_{v_0,\ldots,v_{d'}}\left[\mathcal{T}(v_{d'}) \cdot \sum_{j=1}^{n} a_j \cdot \boldsymbol{v}_j^{(d')}\right] \le \mathop{\mathbb{E}}_{v_0,\ldots,v_{d'}}\left[\mathcal{T}(v_{d'}) \cdot |V|\right], \quad (6)$$

where $V = \sum_{j=1}^{n} a_j \cdot \boldsymbol{v}_j^{(d')}$. Hence by Lemma 28, we have $(6) \le d \cdot \mathbb{E}[\mathcal{T}(v_{d'})] = p \cdot d$.

Second Bound. By Lemma 27, we construct a $2k$-clean parity decision tree $\mathcal{T}'$ of depth $D \le 2d \cdot k$ equivalent to $\mathcal{T}$, where $k = \Theta(\log(1/p))$. Let $U = \sum_{j=1}^{n} a_j \cdot \boldsymbol{u}_j^{(D')}$. Then we have

$$\sum_{j=1}^{n} \left|\widehat{\mathcal{T}}(j)\right| = \sum_{j=1}^{n} \left|\widehat{\mathcal{T}'}(j)\right| = \mathop{\mathbb{E}}_{u_0,\ldots,u_{D'}}\left[\mathcal{T}'(u_{D'}) \cdot \sum_{j=1}^{n} a_j \cdot \boldsymbol{u}_j^{(D')}\right] \le \mathop{\mathbb{E}}_{u_0,\ldots,u_{D'}}\left[\mathcal{T}'(u_{D'}) \cdot |U|\right]. \quad (7)$$

Lemma 30 implies that for all $\varepsilon > 0$, $\mathbf{Pr}[|U| \ge R(\varepsilon)] \le \varepsilon$ where

$$R(\varepsilon) = R(D,d,k,\varepsilon) = O\left(\sqrt{dk \cdot \left(\frac{1}{\varepsilon}\right)^{\frac{1}{k}} \cdot \log\left(\frac{1}{\varepsilon}\right)}\right).$$

For integer $i \ge 1$, let $I_i = \left[R\left(p/2^i\right), R\left(p/2^{i+1}\right)\right]$ and $I_0 = [0, R(p/2)]$ be intervals. Then for each $i \ge 1$, $\mathbf{Pr}[|U| \in I_i] \le p/2^i$. We also know that $\mathbb{E}_{u_0,\ldots,u_{D'}}[\mathcal{T}'(u_{D'})] \le p$. Thus,

$$\begin{aligned}
(7) &= \mathop{\mathbb{E}}_{u_0,\ldots,u_{D'}}\left[\mathcal{T}'(u_{D'}) \cdot |U| \cdot \sum_{i=0}^{+\infty} \mathbb{1}_{|U|\in I_i}\right] \\
&\le R\left(\frac{p}{2}\right) \cdot \mathop{\mathbb{E}}_{u_0,\ldots,u_{D'}}\left[\mathcal{T}'(u_{D'})\right] + \sum_{i=1}^{+\infty} R\left(\frac{p}{2^{i+1}}\right) \cdot \mathop{\mathbb{E}}_{u_0,\ldots,u_{D'}}\left[\mathbb{1}_{|U|\in I_i}\right] \\
&\le \sum_{i=0}^{+\infty} R\left(\frac{p}{2^{i+1}}\right) \cdot \frac{p}{2^i} \\
&= \sum_{i=0}^{+\infty} O\left(p \cdot \sqrt{dk \cdot \left(\frac{2^{i+1}}{p}\right)^{\frac{1}{k}} \cdot \left(\log\left(\frac{1}{p}\right) + i + 1\right)}\right) \cdot \frac{1}{2^i} \\
&= O\left(p \cdot \sqrt{dk \cdot \log\left(\frac{1}{p}\right)}\right) = O\left(p \cdot \sqrt{d} \cdot \log\left(\frac{1}{p}\right)\right).
\end{aligned}$$

◀

[9] If $p < 2^{-d}$, then $p = 0$ and $\mathcal{T} \equiv 0$. If $p > 1/2$, we can consider $\widetilde{\mathcal{T}} = 1 - \mathcal{T}$ by symmetry.

5.2 Level-ℓ Bound

Now we turn to the general levels.

▶ **Lemma 33.** *There exists a universal constant $\tau \geq 1$ such that the following holds. Let $\ell \geq 1$ be an integer. Let $\mathcal{T}\colon \{\pm 1\}^n \to \{0,1\}$ be a depth-D $2k$-clean parity decision tree where $k \geq 4 \cdot \ell$ and $n \geq \max\{\tau, k, D\}$ and any root-to-leaf path has at most d nodes that are $2k$-clean.*

Let $v_0, \ldots, v_{D'}$ be a random root-to-leaf path. Define $\boldsymbol{v}^{(0)}, \ldots, \boldsymbol{v}^{(D')} \in \{-1,0,+1\}^n$ by setting $\boldsymbol{v}^{(i)}_j = \widehat{\mathcal{P}_{v_i}}(j)$ for each $0 \leq i \leq D'$ and $j \in [n]$. Extend $\boldsymbol{v}^{(D'+1)} = \cdots = \boldsymbol{v}^{(D)}$ to equal $\boldsymbol{v}^{(D')}$. Then for any sequence $a_S \in \{-1,0,1\}, S \in \binom{[n]}{\ell}$, any $\varepsilon \leq 1/2$ and $t \in \{0,\ldots,\ell\}$, we have

$$\mathbf{Pr}\left[\exists t' \in \{0,\ldots,t\}, \exists T \in \binom{[n]}{\ell - t'}, \exists i \in [D], \left|\sum_{S \subseteq \overline{T}, |S| = t'} a_{S \cup T} \cdot \boldsymbol{v}^{(i)}_S\right| \geq M(D,d,k,\ell,t',\varepsilon)\right] \leq \varepsilon \cdot t,$$

where we recall that $\boldsymbol{v}^{(i)}_S = \prod_{j \in S} \boldsymbol{v}^{(i)}_j$ and where

$$M(D,d,k,\ell,t',\varepsilon) = \left(\tau \cdot (D + dk) \cdot \left(\frac{n^\ell}{\varepsilon}\right)^{\frac{6}{k}} \log\left(\frac{n^\ell}{\varepsilon}\right)\right)^{t'/2}.$$

Proof. We prove the bound by induction on $t = 0,1,\ldots,\ell$ and show $\tau = 10^4$ suffices. The base case $t = 0$ is trivial, since for any fixed T and i, we always have $\left|a_T \cdot \boldsymbol{v}^{(i)}_\emptyset\right| \leq 1 = M(D,d,k,\ell,0,\varepsilon)$.

Now we focus on the case where $1 \leq t \leq \ell$. For each $0 \leq i \leq D$ and $T \in \binom{[n]}{\ell - t}$, let

$$X^{(i)}_T = \sum_{S \subseteq \overline{T}, |S| = t} a_{S \cup T} \cdot \boldsymbol{v}^{(i)}_S.$$

For $1 \leq i \leq D'$, we have

$$\begin{aligned}
X^{(i)}_T - X^{(i-1)}_T &= \sum_{S \subseteq \overline{T}, |S| = t, S \cap J(v_{i-1}) \neq \emptyset} a_{S \cup T} \cdot \boldsymbol{v}^{(i)}_S \\
&= \sum_{r=1}^{t} \sum_{\substack{U \subseteq J(v_{i-1}) \cap \overline{T}, \\ |U| = r}} \boldsymbol{v}^{(i)}_U \sum_{\substack{V \subseteq \overline{T \cup J(v_{i-1})}, \\ |U| + |V| = t}} a_{T \cup U \cup V} \cdot \boldsymbol{v}^{(i)}_V \\
&= \sum_{r=1}^{t} \sum_{\substack{U \subseteq J(v_{i-1}) \cap \overline{T}, \\ |U| = r}} \boldsymbol{v}^{(i)}_U \sum_{\substack{V \subseteq \overline{T \cup J(v_{i-1})}, \\ |U| + |V| = t}} a_{T \cup U \cup V} \cdot \boldsymbol{v}^{(i-1)}_V \\
&\qquad\qquad \text{(since } \boldsymbol{v}^{(i)}_j = \boldsymbol{v}^{(i-1)}_j \text{ for all } j \notin J(v_{i-1})) \\
&= \sum_{r=1}^{t} \underbrace{\sum_{\substack{U \subseteq J(v_{i-1}) \cap \overline{T}, \\ |U| = r}} \boldsymbol{v}^{(i)}_U \sum_{\substack{V \subseteq \overline{T \cup U}, \\ |U| + |V| = t}} a_{T \cup U \cup V} \cdot \boldsymbol{v}^{(i-1)}_V}_{A(T,r,i)}. \\
&\qquad\qquad \text{(since } \boldsymbol{v}^{(i-1)}_j = 0 \text{ for all } j \in J(v_{i-1}))
\end{aligned}$$

Observe that conditioning on v_{i-1},

- if r is an even number, then $A(T,r,i)$ is a fixed value independent of $\boldsymbol{v}^{(i)}$;
- if r is an odd number, then $A(T,r,i)$ is an unbiased coin with magnitude independent of $\boldsymbol{v}^{(i)}$.

Therefore, trying to apply Lemma 17, we write $X_T^{(i)} - X_T^{(i-1)} = \mu_T^{(i)} + \Delta_T^{(i)} \cdot z_T^{(i)}$, where $z_T^{(1)}, \dots, z_T^{(D)}$ are independent unbiased coins in $\{\pm 1\}$ and $\mu_T^{(i)} = \Delta_T^{(i)} = 0$ for $D' < i \le D$ and

$$\mu_T^{(i)} = \sum_{\substack{r=2,\\ \text{even}}}^{t} A(T,r,i) \quad \text{and} \quad \Delta_T^{(i)} = \left| \sum_{\substack{r=1,\\ \text{odd}}}^{t} A(T,r,i) \right| \quad \text{for } 1 \le i \le D'. \tag{8}$$

First Bound on $A(T,r,i)$. Let $\mathcal{E}_1$ be the following event:

$$\mathcal{E}_1 = \text{“ } \exists \widehat{t} \in \{0, \dots, t-1\}, \exists T' \in \binom{[n]}{\ell - \widehat{t}}, \exists i' \in [D], \left| X_{T'}^{(i')} \right| \ge M\left(D, k, \ell, \widehat{t}, \varepsilon\right) \text{”}.$$

By the induction hypothesis, we have

$$\mathbf{Pr}\left[\mathcal{E}_1\right] \le (t-1) \cdot \varepsilon. \tag{9}$$

We first derive a simple bound, that will be effective for small values of $|J(v_{i-1})|$.

▷ Claim 34. When $\mathcal{E}_1$ does not happen, $|A(T,r,i)| \le |J(v_{i-1})|^r \cdot M(D,d,k,\ell,t-r,\varepsilon)$ holds for all $r \in [t], i \in [D], T \in \binom{[n]}{\ell-t}$.

Proof. Since $\mathcal{E}_1$ does not happen, by union bound we have

$$\begin{aligned} |A(T,r,i)| &= \left| \sum_{\substack{U \subseteq J(v_{i-1}) \cap \overline{T},\\ |U|=r}} \boldsymbol{v}_U^{(i)} \sum_{\substack{V \subseteq \overline{T \cup U},\\ |U|+|V|=t}} a_{T \cup U \cup V} \cdot \boldsymbol{v}_V^{(i-1)} \right| \le |J(v_{i-1})|^r \max_{U \subseteq \overline{T}, |U|=r} \left| X_{T \cup U}^{(i-1)} \right| \\ &\le |J(v_{i-1})|^r \cdot M(D,d,k,\ell,t-r,\varepsilon). \end{aligned}$$ ◁

Second Bound on $A(T,r,i)$. The second bound requires a more refined decomposition on $A(T,r,i)$.

Assume that $c(i-1)$ is the index of $C(v_{i-1})$ in $v_0, \dots, v_{D'}$, i.e., $v_{c(i-1)} = C(v_{i-1})$. This means that $v_{c(i-1)}$ is the closest ancestor to v_{i-1} that is $2k$-clean. Then define

$$L(v_{i-1}) = \bigcup_{c(i-1) \le i' < i-1} J(v_{i'}).$$

The elements of $L(v_{i-1})$ are precisely the coordinates fixed by the queries from $Q_{v_{c(i-1)}}$ to $Q_{v_{i-1}}$, excluding the latter. Since $\mathcal{T}_{C(v_{i-1})}$ makes non-adaptive queries before (and possibly even after) reaching v_i, $L(v_{i-1})$ and $J(v_{i-1})$ depend only on $C(v_{i-1})$ and i. We now expand $A(T,r,i)$ by also grouping terms based on the number of coordinates in $L(v_{i-1})$ as follows:

$$
\begin{aligned}
A(T,r,i) &= \sum_{\substack{U\subseteq J(v_{i-1})\cap\overline{T},\\ |U|=r}} \boldsymbol{v}_U^{(i)} \sum_{\substack{V\subseteq\overline{T\cup U},\\ |U|+|V|=t}} a_{T\cup U\cup V}\cdot \boldsymbol{v}_V^{(i-1)} \\
&= \sum_{r'=0}^{t-r} \sum_{\substack{U\subseteq J(v_{i-1})\cap\overline{T},\\ |U|=r}} \boldsymbol{v}_U^{(i)} \sum_{\substack{W\subseteq L(v_{i-1})\cap\overline{T},\\ |W|=r'}} \boldsymbol{v}_W^{(i-1)} \sum_{\substack{W'\subseteq\overline{T\cup U\cup L(v_{i-1})}\\ |W'|=t-r-r'}} a_{T\cup U\cup W\cup W'}\cdot \boldsymbol{v}_{W'}^{(i-1)} \\
&= \sum_{r'=0}^{t-r} \sum_{\substack{U\subseteq J(v_{i-1})\cap\overline{T},\\ |U|=r}} \boldsymbol{v}_U^{(i)} \sum_{\substack{W\subseteq L(v_{i-1})\cap\overline{T},\\ |W|=r'}} \boldsymbol{v}_W^{(i-1)} \sum_{\substack{W'\subseteq\overline{T\cup U\cup L(v_{i-1})}\\ |W'|=t-r-r'}} a_{T\cup U\cup W\cup W'}\cdot \boldsymbol{v}_{W'}^{c(i-1)} \\
&\qquad\qquad (\text{since } \boldsymbol{v}_j^{(i-1)} = \boldsymbol{v}_j^{c(i-1)} \text{ for all } j\notin L(v_{i-1})) \\
&= \sum_{r'=0}^{t-r} \sum_{\substack{U\subseteq J(v_{i-1})\cap\overline{T},\\ |U|=r}} \boldsymbol{v}_U^{(i)} \sum_{\substack{W\subseteq L(v_{i-1})\cap\overline{T},\\ |W|=r'}} \boldsymbol{v}_W^{(i-1)} \sum_{\substack{W'\subseteq\overline{T\cup U\cup W}\\ |W'|=t-r-r'}} a_{T\cup U\cup W\cup W'}\cdot \boldsymbol{v}_{W'}^{c(i-1)} \\
&\qquad\qquad (\text{since } \boldsymbol{v}_j^{c(i-1)} = 0 \text{ for all } j\in L(v_{i-1})) \\
&= \sum_{r'=0}^{t-r} \underbrace{\sum_{\substack{U\subseteq J(v_{i-1})\cap\overline{T},\\ |U|=r}} \boldsymbol{v}_U^{(i)} \sum_{\substack{W\subseteq L(v_{i-1})\cap\overline{T},\\ |W|=r'}} \boldsymbol{v}_W^{(i-1)} \cdot X_{T\cup U\cup W}^{c(i-1)}}_{\Gamma_T^{(i)}(r,r')}.
\end{aligned}
$$

Since $C(v_{i-1})$ is $2k$-clean, by Fact 19, the collection of random variables

$$
\left\{\boldsymbol{v}_j^{(i)} \,\middle|\, j\in J(v_{i-1})\right\} \cup \left\{\boldsymbol{v}_j^{(i-1)} \,\middle|\, j\in L(v_{i-1})\right\}
$$

is $2k$-wise independent conditioning on $C(v_{i-1})$. Note that $\Gamma_T^{(i)}(r,r')$ is a polynomial of degree at most $r+r'\le\ell<k$, that $\mathbb{E}\left[\Gamma_T^{(i)}(r,r')\,\middle|\,C(v_{i-1})\right]=0$, and

$$
\begin{aligned}
\sigma_T^2(r,r',C(v_{i-1}),i) &:= \mathbb{E}\left[\left(\Gamma_T^{(i)}(r,r')\right)^2 \,\middle|\, C(v_{i-1})\right] \\
&= \sum_{\substack{U\subseteq J(v_{i-1})\cap\overline{T},\\ |U|=r}} \sum_{\substack{W\subseteq L(v_{i-1})\cap\overline{T},\\ |W|=r'}} \left(X_{T\cup U\cup W}^{c(i-1)}\right)^2 \\
&\le \left(|J(v_{i-1})|\right)^r \left(|L(v_{i-1})|\right)^{r'} \left(\max_{|T'|=r+r'+\ell-t,\, i'\in[D]} \left|X_{T'}^{(i')}\right|\right)^2 \\
&\le \left(|J(v_{i-1})|\right)^r D^{r'} \left(\max_{|T'|=r+r'+\ell-t,\, i'\in[D]} \left|X_{T'}^{(i')}\right|\right)^2. \\
&\qquad\qquad (\text{since } |L(v_{i-1})|\le D \text{ by Fact 25})
\end{aligned}
$$

We also have the following claim, the proof of which follows from Lemma 15 applied to the low degree polynomial $\Gamma_T^{(i)}$. The proof is deferred to Appendix C.

▷ Claim 35. $\mathbf{Pr}\left[\mathcal{E}_2\right]\le\varepsilon/3$, where $\mathcal{E}_2$ is the following event: $\exists T\in\binom{[n]}{\ell-t}, i, r, r'$, such that

$$
\left|\Gamma_T^{(i)}(r,r')\right| \ge \left(100\min\left\{k,\log\left(\frac{n^\ell}{\varepsilon}\right)\right\}\cdot\left(\frac{n^\ell}{\varepsilon}\right)^{\frac{6}{k}}\right)^{\frac{r+r'}{2}} \cdot\sigma_T(r,r',C(v_{i-1}),i).
$$

On the other hand, when $\mathcal{E}_1 \vee \mathcal{E}_2$ does not happen, the following calculation holds for all $T \in \binom{[n]}{\ell-t}$, $i \in [D']$, $r \in [t]$, $0 \le r' \le t-r$:

$$\begin{aligned}
&\left|\Gamma_T^{(i)}(r,r')\right| \\
&\le M\left(D,k,\ell,t-r-r',\varepsilon\right) \cdot \sqrt{\left(100 \min\left\{k, \log\left(\frac{n^\ell}{\varepsilon}\right)\right\} \cdot \left(\frac{n^\ell}{\varepsilon}\right)^{\frac{6}{k}}\right)^{r+r'} \left(|J(v_{i-1})|\right)^r \cdot D^{r'}} \\
&\le M\left(D,k,\ell,t-r-r',\varepsilon\right) \cdot \sqrt{\left(100 \cdot \left(\frac{n^\ell}{\varepsilon}\right)^{\frac{6}{k}}\right)^{r+r'} \left(|J(v_{i-1})| \cdot k\right)^r \cdot \left(D \cdot \log\left(\frac{n^\ell}{\varepsilon}\right)\right)^{r'}} \\
&= \sqrt{\left(\tau(D+dk)\left(\frac{n^\ell}{\varepsilon}\right)^{\frac{6}{k}} \log\left(\frac{n^\ell}{\varepsilon}\right)\right)^{t-r-r'} \left(100\left(\frac{n^\ell}{\varepsilon}\right)^{\frac{6}{k}}\right)^{r+r'} \left(|J(v_{i-1})| \cdot k\right)^r \left(D \cdot \log\left(\frac{n^\ell}{\varepsilon}\right)\right)^{r'}} \\
&\le \sqrt{\left(\tau(D+dk)\left(\frac{n^\ell}{\varepsilon}\right)^{\frac{6}{k}} \log\left(\frac{n^\ell}{\varepsilon}\right)\right)^{t} \left(\frac{100}{\tau}\right)^{r+r'} \left(\frac{|J(v_{i-1})|}{d \cdot \log(n^\ell/\varepsilon)}\right)^r} \\
&\le \sqrt{\left(\tau(D+dk)\left(\frac{n^\ell}{\varepsilon}\right)^{\frac{6}{k}} \log\left(\frac{n^\ell}{\varepsilon}\right)\right)^{t} \left(\frac{200}{\tau}\right)^{r+r'} \left(\frac{|J(v_{i-1})|}{2d}\right)^r \frac{1}{\log(n^\ell/\varepsilon)}} \\
&= M(D,d,k,\ell,t,\varepsilon) \cdot \sqrt{\left(\frac{200}{\tau}\right)^{r+r'} \left(\frac{|J(v_{i-1})|}{2d}\right)^r \frac{1}{\log(n^\ell/\varepsilon)}}.
\end{aligned}$$

Hence we have a second bound on $A(T,r,i)$.

▷ Claim 36. When $\mathcal{E}_1 \vee \mathcal{E}_2$ does not happen, the following holds for all $r \in [t], i \in [D], T \in \binom{[n]}{\ell-t}$:

$$|A(T,r,i)| \le \frac{M(D,d,k,\ell,t,\varepsilon)}{\sqrt{\log\left(n^\ell/\varepsilon\right)}} \cdot \sqrt{\left(\frac{800}{\tau}\right)^r \left(\frac{|J(v_{i-1})|}{2d}\right)^r}.$$

Proof. Since $\mathcal{E}_1 \vee \mathcal{E}_2$ does not happen, by union bound and noticing $\tau \ge 800$ we have

$$\begin{aligned}
&|A(T,r,i)| \\
&\le \sum_{r'=0}^{t-r} \left|\Gamma_T^{(i)}(r,r')\right| \le \frac{M(D,d,k,\ell,t,\varepsilon)}{\sqrt{\log\left(n^\ell/\varepsilon\right)}} \cdot \sqrt{\left(\frac{200}{\tau}\right)^r \left(\frac{|J(v_{i-1})|}{2d}\right)^r} \cdot \sum_{r'=0}^{+\infty} \left(\frac{200}{\tau}\right)^{r'/2} \\
&\le \frac{M(D,d,k,\ell,t,\varepsilon)}{\sqrt{\log\left(n^\ell/\varepsilon\right)}} \cdot \sqrt{\left(\frac{800}{\tau}\right)^r \left(\frac{|J(v_{i-1})|}{2d}\right)^r}.
\end{aligned}$$

◁

Final Bound on $\mu_T^{(i)}$ and $\delta_T^{(i)}$. Combining Claim 34 and Claim 36, if $\mathcal{E}_1 \vee \mathcal{E}_2$ does not happen we have

$$|A(T,r,i)| \le M(D,d,k,\ell,t-r,\varepsilon) + \frac{M(D,d,k,\ell,t,\varepsilon)}{\sqrt{\log\left(n^\ell/\varepsilon\right)}} \cdot \sqrt{\left(\frac{800}{\tau}\right)^r \left(\frac{|J(v_{i-1})|}{2d}\right)^r} \cdot 1_{|J(v_{i-1})|>1} \quad (10)$$

To see this, if $|J(v_{i-1})| \le 1$, we use the bound from Claim 34 as the first term in (10). Otherwise $|J(v_{i-1})| > 1$, in which case we use the bound from Claim 36 as the second term in (10).

By Corollary 26, we can now bound $\sum_{i=1}^{D} \left|\mu_T^{(i)}\right|$ and $\sum_{i=1}^{D} \left|\Delta_T^{(i)}\right|^2$ as Claim 37. Its proof is deferred in Appendix D.

▷ Claim 37. When $\mathcal{E}_1 \vee \mathcal{E}_2$ does not happen, $\sum_{i=1}^{D} \left|\mu_T^{(i)}\right| \le R$ and $\sum_{i=1}^{D} \left|\Delta_T^{(i)}\right|^2 \le R^2$ hold for all $T \in \binom{[n]}{\ell-t}$, where

$$R = \frac{M(D,d,k,\ell,t,\varepsilon)}{5 \cdot \sqrt{\log\left(n^\ell/\varepsilon\right)}}. \tag{11}$$

Complete Induction. Let $\beta = \sqrt{2 \cdot \log\left(n^\ell/\varepsilon\right)} \ge 1$ and observe that

$$\begin{aligned} R + \beta \cdot \sqrt{2} \cdot R &\le \beta \cdot 2\sqrt{2} \cdot R && \text{(due to } \beta \ge 1\text{)} \\ &= \frac{2\sqrt{2} \cdot \sqrt{2 \cdot \log\left(n^\ell/\varepsilon\right)}}{5 \cdot \sqrt{\log\left(n^\ell/\varepsilon\right)}} \cdot M(D,d,k,\ell,t,\varepsilon) && \text{(due to (11))} \\ &\le M(D,d,k,\ell,t,\varepsilon). \end{aligned}$$

Then we have

$$\begin{aligned} &\mathbf{Pr}\left[\exists t' \in \{0,\dots,t\}, \exists T' \in \binom{[n]}{\ell - t'}, \exists i \in [D],\ \left|X_T^{(i)}\right| \ge M\left(D,d,k,\ell,t',\varepsilon\right)\right] \\ &= \mathbf{Pr}\left[\mathcal{E}_1 \bigvee \left(\exists T \in \binom{[n]}{\ell - t}, \exists i \in [D],\ \left|X_T^{(i)}\right| \ge M\left(D,d,k,\ell,t,\varepsilon\right)\right)\right] \\ &\le \mathbf{Pr}\left[(\mathcal{E}_1 \vee \mathcal{E}_2) \bigvee \left(\exists T \in \binom{[n]}{\ell - t}, \exists i \in [D],\ \left|X_T^{(i)}\right| \ge R + \beta \cdot \sqrt{2} \cdot R\right)\right] \\ &\le (t-1) \cdot \varepsilon + \frac{\varepsilon}{3} + 2n^{\ell - t} \cdot e^{-\beta^2/2} \qquad \text{(due to (9), Claim 35, Lemma 17, and Claim 37)} \\ &\le (t-1) \cdot \varepsilon + \frac{\varepsilon}{3} + \frac{1}{3} \cdot n^\ell \cdot e^{-\beta^2/2} \\ &\le t \cdot \varepsilon. \end{aligned}$$

◀

Before we prove the complete level-ℓ bound for parity decision trees, we first prove a simple bound for the number of vectors with a given weight in a subspace.

▶ **Lemma 38.** *Let $\ell \ge 1$ be an integer and $\mathcal{S}$ be a subspace of rank at most d. Let $U = \{S \mid |S| = \ell, S \in \mathcal{S}\}$, then $|U| \le \min\left\{\binom{d \cdot \ell}{\ell}, 2^d - 1\right\}$.*

Proof. Let $\{S_1, \dots, S_{d'}\}$ be a maximal set of independent vectors in U. Then $d' \le d$ and $|S_i| = \ell$ holds for all $i \in [d']$. Since $U \subseteq \mathsf{Span}\langle S_1, \dots, S_{d'}\rangle$ and $\emptyset \notin U$, we have

$$|U| \le |\mathsf{Span}\langle S_1, \dots, S_{d'}\rangle| - 1 = 2^{d'} - 1 \le 2^d - 1.$$

On the other hand, observe that $U \subseteq \binom{S_1 \cup \dots \cup S_{d'}}{\ell}$, hence we also have

$$|U| \le \left|\binom{S_1 \cup \dots \cup S_{d'}}{\ell}\right| \le \binom{d' \cdot \ell}{\ell} \le \binom{d \cdot \ell}{\ell}.$$

◀

We remark that in Lemma 38, it is conjectured the bound should be $\binom{d+1}{\ell}$ when $d \ge 2 \cdot \ell$ [19, 6].

▶ **Theorem 39.** *Let $\ell \ge 1$ be an integer. Let $\mathcal{T} \colon \{\pm 1\}^n \to \{0,1\}$ be a depth-d parity decision tree where $n \ge \max\{d, \ell\}$. Let $p = \mathbf{Pr}\left[\mathcal{T}(x) = 1\right] \ge 2^{-d}$.*[10] *Then we have*

$$\sum_{S \subseteq [n] : |S| = \ell} \left|\widehat{\mathcal{T}}(S)\right| \le p \cdot \min\left\{\binom{d \cdot \ell}{\ell}, 2^d - 1,\ O\left(\sqrt{d} \cdot \log\left(\frac{n^\ell}{p}\right)\right)^\ell\right\} = O\left(\sqrt{d} \cdot \ell \cdot \log(n)\right)^\ell.$$

[10] If $p < 2^{-d}$, then $p = 0$ and $\mathcal{T} \equiv 0$.

Proof. For any $S \in \binom{[n]}{\ell}$, let $a_S = \mathsf{sgn}\left(\widehat{\mathcal{T}}(S)\right)$. Now we prove the bounds separately.

First Two Bounds. Let $v_0, \ldots, v_{d'}$ be a random root-to-leaf path. Then by the definition of $\widehat{\mathcal{P}_v}$ and $\mathcal{S}_v$ and Fact 19, we have

$$\begin{aligned}\sum_S \left|\widehat{\mathcal{T}}(S)\right| &= \sum_S a_S \cdot \widehat{\mathcal{T}}(S) = \mathop{\mathbb{E}}_{v_0,\ldots,v_{d'}}\left[\mathcal{T}(v_{d'}) \cdot \sum_S a_S \cdot \widehat{\mathcal{P}_{v_{d'}}}(S)\right] \\ &\le \mathop{\mathbb{E}}_{v_0,\ldots,v_{d'}}\left[\mathcal{T}(v_{d'}) \cdot \sum_S \left|\widehat{\mathcal{P}_{v_{d'}}}(S)\right|\right] = \mathop{\mathbb{E}}_{v_0,\ldots,v_{d'}}\left[\mathcal{T}(v_{d'}) \cdot |V|\right], \qquad (12)\end{aligned}$$

where $a_S = \mathsf{sgn}\left(\widehat{\mathcal{T}}(S)\right)$ and $V = \left\{S \in \binom{[n]}{\ell} \,\middle|\, S \in \mathcal{S}_{v_{d'}}\right\}$. Note that

$$\mathsf{rank}\left(\mathcal{S}_{v_{d'}}\right) = \mathsf{rank}\left(\mathsf{Span}\left\langle Q_{v_0}, \ldots, Q_{v_{d'-1}}\right\rangle\right) \le d' \le d.$$

Hence by Lemma 38, we have $(12) \le \min\left\{\binom{d\cdot\ell}{\ell}, 2^d - 1\right\} \cdot \mathbb{E}\left[\mathcal{T}(v_{d'})\right] = p \cdot \min\left\{\binom{d\cdot\ell}{\ell}, 2^d - 1\right\}$.

Third Bound. By Lemma 27, we construct a $2k$-clean parity decision tree $\mathcal{T}'$ of depth $D \le 2d \cdot k$ equivalent to $\mathcal{T}$, where $k = \Theta\left(\log\left(n^\ell/p\right)\right) \ge 4 \cdot \ell$. We also add dummy variables to make sure $n' = \max\{\tau, k, 6D, n\}$, where $\mathcal{T}'$ has n' inputs and τ is the universal constant in Lemma 33.

Let $u_0, \ldots, u_{D'}$ be a random root-to-leaf path in $\mathcal{T}'$. Define $\boldsymbol{u}^{(0)}, \ldots, \boldsymbol{u}^{(D')} \in \{-1, 0, +1\}^n$ by setting $\boldsymbol{u}^{(i)}_j = \widehat{\mathcal{P}_{u_i}}(j)$ for each $0 \le i \le D'$ and $j \in [n]$. Then extend $\boldsymbol{u}^{(D'+1)} = \boldsymbol{u}^{(D'+2)} = \cdots = \boldsymbol{u}^{(D)}$ to equal $\boldsymbol{u}^{(D')}$. By Lemma 23, we have

$$\sum_S \left|\widehat{\mathcal{T}}(S)\right| = \sum_S \left|\widehat{\mathcal{T}'}(S)\right| = \mathop{\mathbb{E}}_{u_0,\ldots,u_{D'}}\left[\mathcal{T}(u_{D'}) \cdot \sum_S a_S \cdot \boldsymbol{u}_S^{(D)}\right] \le \mathop{\mathbb{E}}_{u_0,\ldots,u_{D'}}\left[\mathcal{T}(u_{D'}) \cdot |U|\right], \quad (13)$$

where $U = \sum_S a_S \cdot \boldsymbol{u}_S^{(D)}$.

Now we apply Lemma 33 with $t = \ell, \varepsilon = \Theta\left(p/d^{\ell/2}\right) \le 1/2$ to obtain the following bound[11]

$$M = M(D, d, k, \ell, \ell, \varepsilon) = \left(O\left(\sqrt{d} \cdot \log\left(\tfrac{n^\ell}{p}\right)\right)\right)^\ell$$

such that $\mathbf{Pr}\left[|U| \ge M\right] \le \ell \cdot \varepsilon$. Then, combining the first bound, we have

$$\begin{aligned}(13) &= \mathbb{E}\left[\mathcal{T}(u_{D'}) \cdot |U| \cdot \left(\mathbb{1}_{|U|<M} + \mathbb{1}_{|U|\ge M}\right)\right] \le M \cdot \mathbb{E}\left[\mathcal{T}(u_{D'})\right] + \ell \cdot \varepsilon \cdot \binom{d \cdot \ell}{\ell} \\ &= p \cdot \left(O\left(\sqrt{d} \cdot \log\left(\tfrac{n^\ell}{p}\right)\right)\right)^\ell,\end{aligned}$$

which is maximized at $p = 1$, hence $(13) = O\left(\sqrt{d} \cdot \ell \cdot \log(n)\right)^\ell$ as desired. ◀

[11] Since $n \ge \max\{\ell, d\}$, we know $k = \Theta\left(\log\left(n^\ell/p\right)\right) = O(n^2)$ and $D \le 2d \cdot k = O(n^3)$. Hence $n' = \max\{\tau, k, 6D, n\} = O(n^3)$. Also $n^\ell/\varepsilon \le n^{O(\ell)}/p$ and by our choice of $k = \Theta\left(\log(n^\ell/p)\right)$ we have $\left(n^\ell/\varepsilon\right)^{6/k} = O(1)$.

6 Fourier Bounds for Noisy Decision Trees

Let $\mathcal{T}$ be a noisy decision tree. By adding queries with zero correlation, we assume without loss of generality each root-to-leaf path in the noisy decision tree is of the same length. Let v be any node of $\mathcal{T}$. We use $\mathcal{P}_v$ to denote the uniform distribution over $\{\pm 1\}^n$ *conditioning* on reaching v. Note that $\mathcal{P}_v$ is always a *product distribution*. As before, for any $S \subseteq [n]$ we define $\widehat{\mathcal{P}_v}(S) = \mathbb{E}_{x\sim\mathcal{P}_v}[x_S]$.

▷ Claim 40. Let $\mathcal{T}\colon \{\pm 1\}^n \to \{0,1\}$ be a cost-d noisy decision tree. Let $v_0, \ldots, v_D$ be any root-to-leaf path in $\mathcal{T}$. Define $\boldsymbol{v}^{(0)}, \ldots, \boldsymbol{v}^{(D)} \in [-1,1]^n$ by setting $\boldsymbol{v}^{(i)}_j = \widehat{\mathcal{P}_{v_i}}(j)$ for each $0 \le i \le D$ and $j \in [n]$. Then for any $i \in \{0, \ldots, D-1\}$, $\boldsymbol{v}^{(i+1)}_{q_{v_i}} - \boldsymbol{v}^{(i)}_{q_{v_i}}$ is a mean-zero random variable with magnitude bounded by $2 \cdot |\gamma_{v_i}|$.

Proof. Fix $i \in \{0, \ldots, D-1\}$. For convenience, let $j = q_{v_i}$, $\gamma = \gamma_{v_i}$, and $\alpha = \boldsymbol{v}^{(i)}_j$. Suppose $|\gamma| = 1$ then $\left|v^{(i+1)}_j - v^{(i)}_j\right| \le 2 = 2 \cdot |\gamma_{v_i}|$ as desired. Now we turn to the case $|\gamma| < 1$.

Note that for the distribution $\mathcal{P}_{v_i}$, the measure of $x_j = 1$ (resp., $x_j = -1$) inputs is $(1+\alpha)/2$ (resp., $(1-\alpha)/2$). The measure of $x_j = 1$ (resp., $x_j = -1$) inputs that follow the edge labeled 1 is $a := (1+\alpha)(1+\gamma)/4$ (resp., $b := (1-\alpha)(1-\gamma)/4$). The total measure of inputs that take the edge labeled 1 is $a + b$ and the resulting node v_{i+1} satisfies $\boldsymbol{v}^{(i+1)}_j = (a-b)/(a+b)$. This implies that

$$\boldsymbol{v}^{(i+1)}_j = \begin{cases} \frac{\alpha+\gamma}{1+\gamma\cdot\alpha} & \text{with probability } \frac{1+\gamma\cdot\alpha}{2}, \\ \frac{\alpha-\gamma}{1-\gamma\cdot\alpha} & \text{with probability } \frac{1-\gamma\cdot\alpha}{2}. \end{cases}$$

The above calculation implies

$$\boldsymbol{v}^{(i+1)}_j - \boldsymbol{v}^{(i)}_j = \begin{cases} \gamma \cdot \frac{1-\alpha^2}{1+\gamma\cdot\alpha} & \text{with probability } \frac{1+\gamma\cdot\alpha}{2}, \\ -\gamma \cdot \frac{1-\alpha^2}{1-\gamma\cdot\alpha} & \text{with probability } \frac{1-\gamma\cdot\alpha}{2}, \end{cases}$$

and thus $\boldsymbol{v}^{(i+1)}_j - \boldsymbol{v}^{(i)}_j$ is a mean-zero random variable. Since $\alpha \in [-1,1]$ and $\gamma \in (-1,1)$, we have

$$\max\left\{\frac{1-\alpha^2}{1-\gamma\cdot\alpha}, \frac{1-\alpha^2}{1+\gamma\cdot\alpha}\right\} \le \frac{1-\alpha^2}{1-|\alpha|} = 1 + |\alpha| \le 2,$$

which implies $\left|\boldsymbol{v}^{(i+1)}_j - \boldsymbol{v}^{(i)}_j\right| \le 2 \cdot |\gamma|$. ◁

We now prove the general Fourier bounds. As before, for any $S \subseteq [n]$, let $\boldsymbol{v}^{(i)}_S$ be $\prod_{j\in S} \boldsymbol{v}^{(i)}_j$.

▶ **Lemma 41.** *There exists a universal constant τ such that the following holds. Let $\ell \ge 1$ be an integer. Let $\mathcal{T}\colon \{\pm 1\}^n \to \{0,1\}$ be a cost-d noisy decision tree.*

Let $v_0, \ldots, v_D$ be a random root-to-leaf path in $\mathcal{T}$. Define $\boldsymbol{v}^{(0)}, \ldots, \boldsymbol{v}^{(D)} \in [-1,1]^n$ by setting $\boldsymbol{v}^{(i)}_j = \widehat{\mathcal{P}_{v_i}}(j)$ for each $0 \le i \le D$ and $j \in [n]$. Then for any sequence $a_S \in \{-1,0,1\}$, $S \in \binom{[n]}{\ell}$, any $\varepsilon \le 1/2$ and $t \in \{0, \ldots, \ell\}$, we have

$$\mathbf{Pr}\left[\exists T \in \binom{[n]}{\ell - t}, \exists i \in [D], \left|\sum_{S \subseteq \overline{T}, |S| = t} a_{S\cup T} \cdot \boldsymbol{v}^{(i)}_S\right| \ge S(d,\ell,t,\varepsilon)\right] \le \varepsilon \cdot t,$$

where $S(d,\ell,0,\varepsilon) = 1$ and

$$S(d,\ell,t,\varepsilon) = \sqrt{(\tau\cdot d)^t \cdot \log\left(\frac{n^{\ell-t}}{\varepsilon}\right) \cdots \log\left(\frac{n^{\ell-1}}{\varepsilon}\right)} \qquad \text{for } t \in [\ell].$$

Proof. We prove the bound by induction on t and show $\tau = 32$ suffices. The base case $t = 0$ is trivial, since for any T of size ℓ and any i, we have $\left|a_T \cdot v_\emptyset^{(i)}\right| \le 1 = S(d, \ell, 0, \varepsilon)$.

Now we focus on the case $1 \le t \le \ell$. For any $T \in \binom{[n]}{\le \ell}$, define $X_T^{(0)}, \dots, X_T^{(D)}$ by $X_T^{(i)} = \sum_{S \subseteq \overline{T}, |S|+|T|=\ell} a_{S\cup T} \cdot \boldsymbol{v}_S^{(i)}$. Define $\delta_T^{(i)}$ for $i \in [D]$ as follows:

$$\begin{aligned}\delta_T^{(i)} = X_T^{(i)} - X_T^{(i-1)} &= \sum_{S\subseteq \overline{T}, |S|=t, S \ni q_{v_{i-1}}} a_{S\cup T} \cdot \left(\boldsymbol{v}_S^{(i)} - \boldsymbol{v}_S^{(i-1)}\right) \\ &= \left(\boldsymbol{v}_{q_{v_{i-1}}}^{(i)} - \boldsymbol{v}_{q_{v_{i-1}}}^{(i-1)}\right) \cdot \sum_{S' \subseteq \overline{T \cup \{q_{v_{i-1}}\}}, |S'|=t-1} a_{S' \cup \{q_{v_{i-1}}\} \cup T} \cdot \boldsymbol{v}_S^{(i-1)} \\ &= \left(\boldsymbol{v}_{q_{v_{i-1}}}^{(i)} - \boldsymbol{v}_{q_{v_{i-1}}}^{(i-1)}\right) \cdot X_{T \cup \{q_{v_{i-1}}\}}^{(i-1)}.\end{aligned}$$

Note that by Claim 40 and conditioning on v_{i-1}, $\delta_T^{(i)}$ is a mean-zero random variable.

The induction hypothesis implies that with all but $\varepsilon \cdot (t-1)$ probability, for all $i \in [D]$ and $T' \in \binom{[n]}{\ell - t + 1}$, we have $\left|X_{T'}^{(i)}\right| \le S(d, \ell, t-1, \varepsilon)$. By Claim 40, we have

$$\left|\delta_T^{(i)}\right| = \left|\boldsymbol{v}_{q_{v_{i-1}}}^{(i)} - \boldsymbol{v}_{q_{v_{i-1}}}^{(i-1)}\right| \cdot \left|X_{T\cup\{q_{v_{i-1}}\}}^{(i-1)}\right| \le 2 \cdot \left|\gamma_{v_{i-1}}\right| \cdot S(d, \ell, t-1, \varepsilon).$$

Denote by $\Delta_T^{(i)} = 2 \cdot \left|\gamma_{v_{i-1}}\right| \cdot S(d, \ell, t-1, \varepsilon)$. We can thus express $X_T^{(i)} = X_T^{(i-1)} + \Delta_T^{(i)} \cdot z_T^{(i)}$ where $\left|z_T^{(i)}\right| \le 1$. Then we apply Lemma 17 to the family of martingales $X_T^{(0)}, \dots, X_T^{(D)}, |T| \in \binom{[n]}{\ell - t}$ with difference sequence $\delta_T^{(i)} = \Delta_T^{(i)} \cdot z_T^{(i)}$ satisfying

$$\sum_{i=1}^{D} \left(\Delta_T^{(i)}\right)^2 = 4 \cdot (S(d, \ell, t-1, \varepsilon))^2 \cdot \sum_{i=1}^{D} \left|\gamma_{v_{i-1}}\right|^2 \le 4d \cdot (S(d, \ell, t-1, \varepsilon))^2 .$$

Hence for any $\beta \ge 0$, we have

$$\mathbf{Pr}\left[\exists T \in \binom{[n]}{\ell - t}, \exists i \in [D], \left|X_T^{(i)}\right| \ge 2\beta \cdot \sqrt{2d} \cdot S(d, \ell, t-1, \varepsilon)\right] \le \varepsilon \cdot (t-1) + 2 \cdot n^{\ell - t} \cdot e^{-\beta^2/2}.$$

Since $\varepsilon \le 1/2$, we can set $\beta = 2 \cdot \sqrt{\log(n^{\ell-t}/\varepsilon)}$ so that $2 \cdot n^{\ell-t} \cdot e^{-\beta^2/2} \le \varepsilon$, which completes the induction by noticing

$$2\beta \cdot \sqrt{2d} \cdot S(d, \ell, t-1, \varepsilon) = \sqrt{32 \cdot d \cdot \log\left(\frac{n^{\ell-t}}{\varepsilon}\right)} \cdot S(d, \ell, t-1, \varepsilon) \le S(d, \ell, t, \varepsilon). \qquad \blacktriangleleft$$

▶ **Theorem 42.** *Let $\ell \ge 1$ and $n \ge \max\{\ell, 2\}$ be integers. Let $\mathcal{T}\colon \{\pm 1\}^n \to \{0, 1\}$ be a cost-d noisy decision tree. Let $p = \mathbf{Pr}[\mathcal{T}(x) = 1] \in (0, 1/2]$.*[12] *Then we have*

$$\sum_{S\subseteq[n], |S|=\ell} \left|\widehat{\mathcal{T}}(S)\right| \le p \cdot O(d)^{\ell/2} \cdot \sqrt{\log\left(\frac{1}{p}\right)\left(\log\left(\frac{n^\ell}{p}\right)\right)^{\ell-1}} = O(d)^{\ell/2} \cdot \sqrt{1 + (\ell \log(n))^{\ell-1}}.$$

Proof. For any $S \in \binom{[n]}{\ell}$, let $a_S = \mathsf{sgn}\left(\widehat{\mathcal{T}}(S)\right)$. Let $v_0, \dots, v_D$ be a random root-to-leaf path in $\mathcal{T}$. Note that

$$\sum_S \left|\widehat{\mathcal{T}}(S)\right| = \sum_S a_S \cdot \widehat{\mathcal{T}}(S) = \mathbb{E}\left[\mathcal{T}(v_D) \cdot \sum_S a_S \cdot \boldsymbol{v}_S^{(D)}\right] \le \mathbb{E}\left[\mathcal{T}(v_D) \cdot |V|\right], \tag{14}$$

[12] If $p > 1/2$, then we can consider $\widetilde{\mathcal{T}} = 1 - \mathcal{T}$ by symmetry.

where $V = \sum_S a_S \cdot_S \boldsymbol{v}_S^{(D)}$. By Lemma 41, we know $\mathbf{Pr}[|V| \geq S(\varepsilon)] \leq \varepsilon \cdot \ell$, where

$$S(\varepsilon) = S(d,\ell,\ell,\varepsilon) = \sqrt{O(d)^\ell \cdot \log\left(\tfrac{n^{\ell-1}}{\varepsilon}\right) \cdots \log\left(\tfrac{n^0}{\varepsilon}\right)} \leq \sqrt{O(d)^\ell \cdot \left(\log\left(\tfrac{n^{\ell-1}}{\varepsilon}\right)\right)^{\ell-1} \log\left(\tfrac{1}{\varepsilon}\right)}.$$

For integer $i \geq 1$, let $I_i = \left[S\left(p/\left(\ell 2^i\right)\right), S\left(p/\left(\ell 2^{i+1}\right)\right)\right]$ and $I_0 = [0, S(p/\ell)]$ be intervals. Then for each $i \geq 1$, $\mathbf{Pr}[|V| \in I_i] \leq p/2^i$. We also know that $\mathbb{E}_{v_0,\ldots,v_D}[\mathcal{T}(v_D)] \leq p$. Thus,

$$\begin{aligned}
(14) &\leq \mathop{\mathbb{E}}_{v_0,\ldots,v_D}\left[\mathcal{T}(v_D) \cdot |V| \cdot \sum_{i=0}^{+\infty} 1_{|V|\in I_i}\right] \\
&\leq S\left(\frac{p}{\ell}\right) \cdot \mathbb{E}[\mathcal{T}(v_D)] + \sum_{i=1}^{+\infty} S\left(\frac{p}{\ell \cdot 2^{i+1}}\right) \cdot \mathbb{E}\left[1_{|V|\in I_i}\right] \\
&\leq \sum_{i=0}^{+\infty} S\left(\frac{p}{\ell \cdot 2^{i+1}}\right) \cdot \frac{p}{2^i} \\
&= \sum_{i=0}^{+\infty} p \cdot \sqrt{O(d)^\ell \cdot \left(\log\left(\tfrac{n^{\ell-1}\cdot \ell}{p}\right) + i + 1\right)^{\ell-1} \cdot \left(\log\left(\tfrac{1}{p}\right) + \log(\ell) + i + 1\right)} \cdot \frac{1}{2^i} \\
&\leq \sum_{i=0}^{+\infty} p \cdot \sqrt{O(d)^\ell \cdot \left(\left(\log\left(\tfrac{n^\ell}{p}\right)\right)^{\ell-1} + (i+1)^{\ell-1}\right) \cdot \left(\log\left(\tfrac{1}{p}\right) + i + 1\right)} \cdot \frac{1}{2^i} \\
&\quad \text{(since } n \geq \ell \text{, and } (x+y)^b \leq 2^b \cdot \left(x^b + y^b\right) \text{ and } \sqrt{x+y} \leq \sqrt{x} + \sqrt{y} \text{ for } x, y, b \geq 0) \\
&\leq p \cdot \sqrt{O(d)^\ell \cdot \log\left(\tfrac{1}{p}\right) \left(\log\left(\tfrac{n^\ell}{p}\right)\right)^{\ell-1}},
\end{aligned}$$

where the last inequality follows from $p \leq 1/2$, $n \geq 2$ and

$$\sum_{i=0}^{+\infty} (i+1)^{\ell/2} \cdot 2^{-i} = O(\ell)^{\ell/2} \leq O(1)^\ell \cdot \ell^{(\ell-1)/2} \leq O(1)^\ell \cdot \left(\log\left(n^\ell/p\right)\right)^{(\ell-1)/2}.$$

Note that $p \cdot (\log(1/p))^k \leq O(k)^k$ for $p \in (0,1)$ and $k \geq 0$, thus

$$\begin{aligned}
p \cdot \sqrt{\log\left(\tfrac{1}{p}\right) \left(\log\left(\tfrac{n^\ell}{p}\right)\right)^{\ell-1}} &= p \cdot \sqrt{\log\left(\tfrac{1}{p}\right) \left(\ell \log(n) + \log\left(\tfrac{1}{p}\right)\right)^{\ell-1}} \\
&\leq O(1)^\ell \cdot \left(\sqrt{(\ell \log(n))^{\ell-1}} + \ell^{\ell/2}\right) \\
&= O(1)^\ell \cdot \sqrt{1 + (\ell \log(n))^{\ell-1}}.
\end{aligned}$$

◀

References

1 Scott Aaronson and Andris Ambainis. Forrelation: A problem that optimally separates quantum from classical computing. *SIAM J. Comput.*, 47(3):982–1038, 2018.

2 Nikhil Bansal and Makrand Sinha. k-forrelation optimally separates quantum and classical query complexity. *Electron. Colloquium Comput. Complex.*, 27:127, 2020.

3 Shalev Ben-David and Eric Blais. A tight composition theorem for the randomized query complexity of partial functions: Extended abstract. In *FOCS*, pages 240–246. IEEE, 2020.

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

5 Aline Bonami. Étude des coefficients de fourier des fonctions de $l^p(g)$. *Annales de l'institut Fourier*, 20(2):335–402, 1970. URL: `http://eudml.org/doc/74019`.

6 Joseph Briggs and Wesley Pegden. Extremal collections of k-uniform vectors. *arXiv preprint*, 2018. arXiv:1801.09609.

7 Arkadev Chattopadhyay, Yuval Filmus, Sajin Koroth, Or Meir, and Toniann Pitassi. Query-to-communication lifting for BPP using inner product. In Christel Baier, Ioannis Chatzigiannakis, Paola Flocchini, and Stefano Leonardi, editors, *46th International Colloquium on Automata, Languages, and Programming, ICALP 2019, July 9-12, 2019, Patras, Greece*, volume 132 of *LIPIcs*, pages 35:1–35:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019. doi:10.4230/LIPIcs.ICALP.2019.35.

8 Eshan Chattopadhyay, Jason Gaitonde, Chin Ho Lee, Shachar Lovett, and Abhishek Shetty. Fractional pseudorandom generators from any fourier level. *CoRR*, abs/2008.01316, 2020. arXiv:2008.01316.

9 Eshan Chattopadhyay, Pooya Hatami, Kaave Hosseini, and Shachar Lovett. Pseudorandom generators from polarizing random walks. *Theory Comput.*, 15:1–26, 2019.

10 Eshan Chattopadhyay, Pooya Hatami, Shachar Lovett, and Avishay Tal. Pseudorandom generators from the second fourier level and applications to AC0 with parity gates. In *ITCS*, volume 124 of *LIPIcs*, pages 22:1–22:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019.

11 Eshan Chattopadhyay, Pooya Hatami, Omer Reingold, and Avishay Tal. Improved pseudorandomness for unordered branching programs through local monotonicity. In *STOC*, pages 363–375. ACM, 2018.

12 Gil Cohen, Noam Peri, and Amnon Ta-Shma. Expander random walks: A fourier-analytic approach. *Electron. Colloquium Comput. Complex.*, 27:163, 2020.

13 Gil Cohen and Igor Shinkar. The complexity of DNF of parities. In *ITCS*, pages 47–58. ACM, 2016.

14 Dmitry Gavinsky. Entangled simultaneity versus classical interactivity in communication complexity. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16, page 877–884, New York, NY, USA, 2016. Association for Computing Machinery. doi:10.1145/2897518.2897545.

15 Uma Girish, Ran Raz, and Avishay Tal. Quantum versus randomized communication complexity, with efficient players. In *ITCS*, volume 185 of *LIPIcs*, pages 54:1–54:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.

16 Uma Girish, Ran Raz, and Wei Zhan. Lower bounds for XOR of forrelations. *Electron. Colloquium Comput. Complex.*, 27:101, 2020.

17 Parikshit Gopalan, Rocco A. Servedio, Avishay Tal, and Avi Wigderson. Degree and sensitivity: tails of two distributions. *Electron. Colloquium Comput. Complex.*, 23:69, 2016.

18 Hamed Hatami, Kaave Hosseini, and Shachar Lovett. Structure of protocols for XOR functions. *SIAM J. Comput.*, 47(1):208–217, 2018.

19 Joshua Brown Kramer. On the most weight w vectors in a dimension k binary code. *Electron. J. Comb.*, 17(1), 2010.

20 Raghav Kulkarni, Youming Qiao, and Xiaoming Sun. On the power of parity queries in boolean decision trees. In *TAMC*, volume 9076 of *Lecture Notes in Computer Science*, pages 99–109. Springer, 2015.

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

22 Chin Ho Lee. Fourier bounds and pseudorandom generators for product tests. In *Computational Complexity Conference*, volume 137 of *LIPIcs*, pages 7:1–7:25. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2019.

23 Nikhil S. Mande and Swagato Sanyal. On parity decision trees for fourier-sparse boolean functions. In *FSTTCS*, volume 182 of *LIPIcs*, pages 29:1–29:16. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

24 Yishay Mansour. An o(nˆ(log log n)) learning algorithm for DNF under the uniform distribution. *J. Comput. Syst. Sci.*, 50(3):543–550, 1995.

25 Ashley Montanaro, Harumichi Nishimura, and Rudy Raymond. Unbounded-error quantum query complexity. *Theor. Comput. Sci.*, 412(35):4619–4628, 2011. doi:10.1016/j.tcs.2011.04.043.

26 Ashley Montanaro and Tobias Osborne. On the communication complexity of XOR functions. *CoRR*, abs/0909.3392, 2009. arXiv:0909.3392.

27 Joseph Naor and Moni Naor. Small-bias probability spaces: Efficient constructions and applications. *SIAM J. Comput.*, 22(4):838–856, 1993.

28 Ryan O'Donnell. Open problems in analysis of boolean functions. *CoRR*, abs/1204.6447, 2012. arXiv:1204.6447.

29 Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

30 Ryan O'Donnell and Rocco A. Servedio. Learning monotone decision trees in polynomial time. *SIAM J. Comput.*, 37(3):827–844, 2007.

31 Ryan O'Donnell, John Wright, Yu Zhao, Xiaorui Sun, and Li-Yang Tan. A composition theorem for parity kill number. In *Computational Complexity Conference*, pages 144–154. IEEE Computer Society, 2014.

32 Omer Reingold, Thomas Steinke, and Salil P. Vadhan. Pseudorandomness for regular branching programs via fourier analysis. In *APPROX-RANDOM*, volume 8096 of *Lecture Notes in Computer Science*, pages 655–670. Springer, 2013.

33 Swagato Sanyal. Fourier sparsity and dimension. *Theory Comput.*, 15:1–13, 2019.

34 Alexander A. Sherstov, Andrey A. Storozhenko, and Pei Wu. An optimal separation of randomized and quantum query complexity. *Electron. Colloquium Comput. Complex.*, 27:128, 2020.

35 Amir Shpilka, Avishay Tal, and Ben lee Volk. On the structure of boolean functions with small spectral norm. *Comput. Complex.*, 26(1):229–273, 2017.

36 Avishay Tal. Tight bounds on the fourier spectrum of AC0. In *Computational Complexity Conference*, volume 79 of *LIPIcs*, pages 15:1–15:31. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017.

37 Avishay Tal. Towards optimal separations between quantum and randomized query complexities. In *FOCS*, pages 228–239. IEEE, 2020.

38 Hing Yin Tsang, Chung Hoi Wong, Ning Xie, and Shengyu Zhang. Fourier sparsity, spectral norm, and the log-rank conjecture. In *FOCS*, pages 658–667. IEEE Computer Society, 2013.

39 Zhiqiang Zhang and Yaoyun Shi. Communication complexities of symmetric XOR functions. *Quantum Inf. Comput.*, 9(3&4):255–263, 2009.

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

A Proof of Corollary 8

▶ **Corollary** (Corollary 8 restated). *Let $\mathcal{T}$ be a parity decision tree of size at most $s > 1$ on n variables. Then,*

$$\forall \ell \in [n] : L_{1,\ell}(f) \leq (\log(s))^{\ell/2} \cdot O(\ell \cdot \log(n))^{1.5\ell}.$$

Proof. We approximate $\mathcal{T}$ with error $\varepsilon = 1/n^\ell$ by another parity decision tree $\mathcal{T}'$ of depth $d = \lceil \log\left(s \cdot n^\ell\right) \rceil$, where we simply replace all nodes of depth d in $\mathcal{T}$ with leaves that return 0. Since there are at most s nodes in $\mathcal{T}$, the probability that a random input would reach one of the nodes of depth d is at most $2^{-d} \cdot s \leq 1/n^\ell$. Hence $\mathbf{Pr}_x\left[\mathcal{T}(x) \neq \mathcal{T}'(x)\right] \leq \varepsilon$. This implies that $\left|\widehat{\mathcal{T}}(S) - \widehat{\mathcal{T}'}(S)\right| \leq \varepsilon$ for any subset $S \subseteq [n]$. Thus,

$$L_{1,\ell}(\mathcal{T}) = \sum_{S:|S|=\ell} \left|\widehat{\mathcal{T}}(S)\right| \leq \sum_{S:|S|=\ell} \left(\left|\widehat{\mathcal{T}'}(S)\right| + \varepsilon\right) \leq L_{1,\ell}(\mathcal{T}') + 1.$$

Since $\mathcal{T}'$ is of depth at most $d = \lceil \log(s) + \ell \cdot \log(n) \rceil = O\left(\log(s) \cdot \ell \cdot \log(n)\right)$, we obtain our bound. ◀

B Proof of Lemma 16

We will use the definition of sub-Gaussian random variables.

▶ **Definition 43** (Sub-Gaussian random variables). *We say a random variable x is Δ-sub-Gaussian if $\mathbb{E}\left[e^{t\cdot x}\right] \le e^{t^2\Delta^2}$ holds for all $t \in \mathbb{R}$.*

Now we prove the following sub-Gaussian adaptive Azuma's inequality.

▶ **Lemma 44** (Sub-Gaussian adaptive Azuma's inequality). *Let $X^{(0)}, \dots, X^{(D)}$ be a martingale with respect to a filtration $\left(\mathcal{F}^{(i)}\right)_{i=0}^{D}$[13] and $\Delta^{(1)}, \dots, \Delta^{(D)}$ be a sequence of magnitudes such that $X^{(0)} = 0$ and $X^{(i)} = X^{(i-1)} + \delta^{(i)}$ for $i \in [D]$, where if conditioning on $\mathcal{F}^{(i-1)}$, $\delta^{(i)}$ is a $\Delta^{(i)}$-sub-Gaussian random variable and $\Delta^{(i)}$ is a fixed value.*

If there exists some constant $U \ge 0$ such that $\sum_{i=1}^{D} \left|\Delta^{(i)}\right|^2 \le U$ always holds, then for any $\beta \ge 0$ we have

$$\mathbf{Pr}\left[\max_{i=0,1,\dots,D} \left|X^{(i)}\right| \ge \beta \cdot \sqrt{2U}\right] \le 2 \cdot e^{-\beta^2/2}.$$

Proof. The bound holds trivially when $\beta = 0$, hence we assume $\beta > 0$ from now on. We construct another martingale $\widehat{X}^{(0)}, \dots, \widehat{X}^{(D)}$ as follows:

$$\widehat{X}^{(i)} = \begin{cases} X^{(i)} & 0 \le i \le d, \\ X^{(d)} & i > d, \end{cases} \quad \text{where} \quad d = \min\,\{D\} \cup \left\{ i \in \{0, 1 \dots, D\} \;\middle|\; \left|X^{(i)}\right| \ge \beta \cdot \sqrt{2U} \right\}.$$

We write $\widehat{\delta}^{(i)} = \widehat{X}^{(i)} - \widehat{X}^{(i-1)}$, then $\widehat{\delta}^{(i)} = \delta^{(i)}$ for all $i \le d$; and $\widehat{\delta}^{(i)} \equiv 0$ for all $i > d$. Let $\widehat{\Delta}^{(i)} = \Delta^{(i)}$ for all $i \le d$; and $\widehat{\Delta}^{(i)} \equiv 0$ for all $i > d$. Thus $\widehat{\delta}^{(i)}$ is $\widehat{\Delta}^{(i)}$-sub-Gaussian given $\mathcal{F}^{(i-1)}$; and

$$\sum_{i=1}^{D} \left|\widehat{\Delta}^{(i)}\right|^2 = \sum_{i=1}^{d} \left|\Delta^{(i)}\right|^2 \le U.$$

Moreover, we have

$$\mathbf{Pr}\left[\max_{i=0,1,\dots,D} \left|X^{(i)}\right| \ge \beta \cdot \sqrt{2U}\right] = \mathbf{Pr}\left[\left|\widehat{X}^{(D)}\right| \ge \beta \cdot \sqrt{2U}\right].$$

Let $t > 0$ be a parameter and we bound $\mathbb{E}\left[e^{t\cdot\widehat{X}^{(D)}}\right]$ as follows

$$\mathbb{E}\left[e^{t\cdot\widehat{X}^{(D)}}\right] = \mathop{\mathbb{E}}_{\mathcal{F}^{(D-1)}}\left[e^{t\cdot\widehat{X}^{(D-1)}} \cdot \mathop{\mathbb{E}}_{\mathcal{F}^{(D)}}\left[e^{t\cdot\left(\widehat{X}^{(D)}-\widehat{X}^{(D-1)}\right)} \,\middle|\, \mathcal{F}^{(D-1)}\right]\right] \tag{15}$$

$$= \mathop{\mathbb{E}}_{\mathcal{F}^{(D-1)}}\left[e^{t\cdot\widehat{X}^{(D-1)}} \cdot \mathop{\mathbb{E}}_{\mathcal{F}^{(D)}}\left[e^{t\cdot\widehat{\delta}^{(D)}} \,\middle|\, \mathcal{F}^{(D-1)}\right]\right] \tag{16}$$

[13] $\mathcal{F}^{(0)} \subseteq \mathcal{F}^{(1)} \subseteq \dots \subseteq \mathcal{F}^{(D)}$ is an increasing sequence of σ-algebra where each $\mathcal{F}^{(i)}$ makes $X^{(0)}, \dots, X^{(i+1)}$ measurable and $\mathbb{E}\left[X^{(i)} \,\middle|\, \mathcal{F}^{(i-1)}\right] = X^{(i-1)}$. Intuitively, the filtration is the history of the martingale.

$$\le \mathop{\mathbb{E}}_{\mathcal{F}^{(D-1)}}\left[e^{t\cdot \widehat{X}^{(D-1)}}\cdot e^{t^2\left(\widehat{\Delta}^{(D)}\right)^2}\right] \qquad \text{(since } \widehat{\delta}^{(D)} \text{ is } \widehat{\Delta}^{(D)}\text{-sub-Gaussian)}$$

$$\le \mathop{\mathbb{E}}_{\mathcal{F}^{(D-1)}}\left[e^{t\cdot \widehat{X}^{(D-1)}}\cdot e^{t^2\left(U-\left(\widehat{\Delta}^{(1)}\right)^2-\cdots-\left(\widehat{\Delta}^{(D-1)}\right)^2\right)}\right]$$

$$\le \mathop{\mathbb{E}}_{\mathcal{F}^{(D-2)}}\left[e^{t\cdot \widehat{X}^{(D-2)}}\cdot e^{t^2\left(U-\left(\widehat{\Delta}^{(1)}\right)^2-\cdots-\left(\widehat{\Delta}^{(D-1)}\right)^2\right)}e^{t^2\left(\widehat{\Delta}^{(D-1)}\right)^2}\right] \qquad \text{(similar to (15) and (16))}$$

$$= \mathop{\mathbb{E}}_{\mathcal{F}^{(D-2)}}\left[e^{t\cdot \widehat{X}^{(D-2)}}\cdot e^{t^2\left(U-\left(\widehat{\Delta}^{(1)}\right)^2-\cdots-\left(\widehat{\Delta}^{(D-2)}\right)^2\right)}\right]$$

$$\le \cdots \le \mathop{\mathbb{E}}_{\mathcal{F}^{(D-k)}}\left[e^{t\cdot \widehat{X}^{(D-k)}}\cdot e^{t^2\left(U-\left(\widehat{\Delta}^{(1)}\right)^2-\cdots-\left(\widehat{\Delta}^{(D-k)}\right)^2\right)}\right] \le \cdots$$

$$\le e^{t^2 U}. \tag{17}$$

Setting $t = \beta/\sqrt{2U}$ implies that

$$\mathbf{Pr}\left[\widehat{X}^{(D)} \ge \beta\cdot\sqrt{2U}\right] \le \frac{\mathbb{E}\left[e^{t\cdot\widehat{X}^{(D)}}\right]}{e^{t\cdot\beta\cdot\sqrt{2U}}} \le \frac{e^{t^2U}}{e^{\beta^2}} = e^{-\beta^2/2}.$$

Similarly we can show $\mathbf{Pr}\left[\widehat{X}^{(D)} \le -\beta\cdot\sqrt{2U}\right] \le e^{-\beta^2/2}$, which completes the proof by a union bound. ◀

For our applications, we need the following fact.

▶ **Fact 45.** *Let x be a mean-zero random variable and assume $|x| \le \Delta$ always holds. Then x is Δ-sub-Gaussian.*

Proof. Note that $e^{t\cdot x}$ is convex for all $t \in \mathbb{R}$. By Jensen's inequality, we have

$$\mathbb{E}\left[e^{t\cdot x}\right] \le \frac{1}{2}\left(e^{-t\Delta}+e^{t\Delta}\right) = \sum_{i=0}^{+\infty}\frac{(t\Delta)^{2i}}{(2i)!} \le \sum_{i=0}^{+\infty}\frac{(t\Delta)^{2i}}{i!} = e^{t^2\Delta^2}.$$ ◀

As a corollary of Lemma 44 and Fact 45, we obtain Lemma 16.

▶ **Corollary** (Lemma 16 restated). *Let $X^{(0)},\ldots,X^{(D)}$ be a martingale and $\Delta^{(1)},\ldots,\Delta^{(D)}$ be a sequence of magnitudes such that $X^{(0)} = 0$ and $X^{(i)} = X^{(i-1)} + \Delta^{(i)}\cdot z^{(i)}$ for $i \in [D]$, where if conditioning on $z^{(1)},\ldots,z^{(i-1)}$,*

(1) *$z^{(i)}$ is a mean-zero random variable and $\left|z^{(i)}\right| \le 1$ always holds;*

(2) *$\Delta^{(i)}$ is a fixed value.*

If there exists some constant $U \ge 0$ such that $\sum_{i=1}^{D}\left|\Delta^{(i)}\right|^2 \le U$ always holds, then for any $\beta \ge 0$ we have

$$\mathbf{Pr}\left[\max_{i=0,1,\ldots,D}\left|X^{(i)}\right| \ge \beta\cdot\sqrt{2U}\right] \le 2\cdot e^{-\beta^2/2}.$$

C Proof of Claim 35

▷ Claim (Claim 35 restated). $\mathbf{Pr}\left[\mathcal{E}_2\right] \le \varepsilon/3$, where $\mathcal{E}_2$ is the following event: $\exists T \in \binom{[n]}{\ell-t}, i, r, r'$, such that

$$\left|\Gamma_T^{(i)}(r,r')\right| \ge \left(100\min\left\{k,\log\left(\frac{n^\ell}{\varepsilon}\right)\right\}\cdot\left(\frac{n^\ell}{\varepsilon}\right)^{\frac{6}{k}}\right)^{\frac{r+r'}{2}}\cdot\sigma_T(r,r',C(v_{i-1}),i).$$

Proof. Let $k' = \min\left\{k, \lceil 6\log\left(n^\ell/\varepsilon\right)\rceil\right\} \le 12\min\left\{k, \log\left(n^\ell/\varepsilon\right)\right\}$. Then $\mathcal{T}$ is also a depth-D $2k'$-clean parity decision tree. Observe that

$$\begin{aligned}
&\mathbf{Pr}\left[\left|\Gamma_T^{(i)}(r,r')\right| \ge \left(\frac{4k'}{\eta^{2/k'}}\right)^{(r+r')/2} \cdot \sigma_T(r,r',C(v_{i-1}),i)\right]\\
&\le \max_{C(v_{i-1})} \mathbf{Pr}\left[\left|\Gamma_T^{(i)}(r,r')\right| \ge \left(\frac{4k'}{\eta^{2/k'}}\right)^{(r+r')/2} \cdot \sigma_T(r,r',C(v_{i-1}),i) \,\middle|\, C(v_{i-1})\right]\\
&\le \underbrace{\frac{(4\cdot k')^{r+r'}}{(2\cdot(r+r'))^{k'}}}_{\le 1} \cdot \underbrace{\eta^{2-\frac{2(r+r')}{k'}}}_{\le \eta}\\
&\qquad\qquad\qquad\qquad \text{(due to the second bound in Lemma 15 and } k \ge 4\cdot\ell \ge 4\cdot(r+r'))\\
&\le \eta.
\end{aligned}$$

Thus by union bound over all $T \in \binom{[n]}{\ell-t}, i \in [D'], r \in [t], 0 \le r' \le t-r$, we have

$$\mathbf{Pr}\left[\exists T,i,r,r',\ \left|\Gamma_T^{(i)}(r,r')\right| \ge \left(\tfrac{4k}{\eta^{2/k}}\right)^{(r+r')/2} \cdot \sigma_T(r,r',C(v_{i-1}),i)\right] \le Dt^2 n^{\ell-t}\cdot\eta \le \tfrac{n^{3\cdot\ell}\cdot\eta}{3},$$

where we use the fact $n \ge \max\{D, 3\cdot t\}$ and $t \ge 1$. By setting $\eta = \varepsilon/n^{3\cdot\ell}$, we have

$$\frac{4k'}{\eta^{2/k'}} = 4k'\left(\frac{n^{3\cdot\ell}}{\varepsilon}\right)^{\frac{2}{k'}} \le 4k'\left(\frac{n^\ell}{\varepsilon}\right)^{\frac{6}{k'}} \le 4\cdot 12\min\left\{k, \log\left(\frac{n^\ell}{\varepsilon}\right)\right\}\cdot 2\left(\frac{n^\ell}{\varepsilon}\right)^{\frac{6}{k}},$$

as desired. ◁

D Proof of Claim 37

We first need the following simple bound on M.

▶ **Lemma 46.** *For any integer $s \ge 1$, we have*

$$\sum_{r=s}^{t} M(D,d,k,\ell,t-r,\varepsilon) \le \frac{2\cdot M(D,d,k,\ell,t,\varepsilon)}{\left(\tau D\cdot\log\left(n^\ell/\varepsilon\right)\right)^{s/2}}.$$

Proof. We simply expand the formula of M as follows:

$$\begin{aligned}
\frac{\sum_{r=s}^{t} M(D,d,k,\ell,t-r,\varepsilon)}{M(D,d,k,\ell,t,\varepsilon)} &= \sum_{r=s}^{t}\left(\tau\cdot(D+dk)\cdot\left(\tfrac{n^\ell}{\varepsilon}\right)^{6/k}\log\left(\tfrac{n^\ell}{\varepsilon}\right)\right)^{-r/2}\\
&\le \sum_{r=s}^{+\infty}\left(\tau\cdot(D+dk)\cdot\left(\tfrac{n^\ell}{\varepsilon}\right)^{6/k}\log\left(\tfrac{n^\ell}{\varepsilon}\right)\right)^{-r/2}\\
&\le 2\cdot\left(\tau\cdot(D+dk)\cdot\left(\tfrac{n^\ell}{\varepsilon}\right)^{6/k}\log\left(\tfrac{n^\ell}{\varepsilon}\right)\right)^{-s/2}\\
&\qquad\qquad\qquad \text{(due to } \tau \ge 4 \text{ and } s \ge 1)\\
&\le 2\cdot\left(\tau D\cdot\log\left(n^\ell/\varepsilon\right)\right)^{-s/2}.
\end{aligned}$$
◀

Now we prove Claim 37.

▷ Claim (Claim 37 restated). When $\mathcal{E}_1 \vee \mathcal{E}_2$ does not happen, $\sum_{i=1}^{D} \left|\mu_T^{(i)}\right| \le R$ and $\sum_{i=1}^{D} \left|\delta_T^{(i)}\right|^2 \le R^2$ hold for all $T \in \binom{[n]}{\ell-t}$, where

$$R = \frac{M(D,d,k,\ell,t,\varepsilon)}{5 \cdot \sqrt{\log(n^\ell/\varepsilon)}}.$$

Proof. We verify for each $T \in \binom{[n]}{\ell-t}$ as follows:

$$\begin{aligned}
&\sum_{i=1}^{D} \left|\mu_T^{(i)}\right| \\
&= \sum_{i=1}^{D'} \left|\mu_T^{(i)}\right| \le \sum_{i=1}^{D'} \sum_{\substack{r=2,\\ \text{even}}}^{t} |A(T,r,i)| && \text{(due to (8))}\\
&\le \sum_{i=1}^{D'} \sum_{\substack{r=2,\\ \text{even}}}^{t} \left(M(D,d,k,\ell,t-r,\varepsilon) + \tfrac{M(D,d,k,\ell,t,\varepsilon)}{\sqrt{\log(n^\ell/\varepsilon)}} \cdot \sqrt{\left(\tfrac{800}{\tau}\right)^r \left(\tfrac{|J(v_{i-1})|}{2d}\right)^r} \cdot 1_{|J(v_{i-1})|>1} \right) \\
& && \text{(due to (10))}\\
&\le \sum_{i=1}^{D'} \sum_{\substack{r=2,\\ \text{even}}}^{t} \left(M(D,d,k,\ell,t-r,\varepsilon) + \tfrac{M(D,d,k,\ell,t,\varepsilon)}{\sqrt{\log(n^\ell/\varepsilon)}} \cdot \left(\tfrac{|J(v_{i-1})|}{2d}\right) \left(\tfrac{800}{\tau}\right)^{r/2} \cdot 1_{|J(v_{i-1})|>1} \right) \\
& && \text{(Since } |J(v_{i-1})| \le 2d \text{ from Corollary 26)}\\
&\le \tfrac{2\cdot M(D,d,k,\ell,t,\varepsilon)}{\tau\cdot\log(n^\ell/\varepsilon)} + \tfrac{1.1\cdot 800\cdot M(D,d,k,\ell,t,\varepsilon)}{\tau\cdot\sqrt{\log(n^\ell/\varepsilon)}} \\
& && \text{(due to Lemma 46 and Corollary 26 and } \tau = 10^4)\\
&\le \tfrac{M(D,d,k,\ell,t,\varepsilon)}{5\cdot\sqrt{\log(n^\ell/\varepsilon)}} = R
\end{aligned}$$

and with similar calculation, we have

$$\begin{aligned}
&\sum_{i=1}^{D} \left|\delta_T^{(i)}\right|^2 \\
&\le \sum_{i=1}^{D'} \left(\sum_{\substack{r=1,\\ \text{odd}}}^{t} \left(M(D,d,k,\ell,t-r,\varepsilon) + \tfrac{M(D,d,k,\ell,t,\varepsilon)}{\sqrt{\log(n^\ell/\varepsilon)}} \cdot \sqrt{\tfrac{|J(v_{i-1})|}{2d}} \left(\tfrac{800}{\tau}\right)^{r/2} \cdot 1_{|J(v_{i-1})|>1} \right) \right)^2 \\
&\le \sum_{i=1}^{D'} \left(\tfrac{2\cdot M(D,d,k,\ell,t,\varepsilon)}{\sqrt{\tau D\cdot\log(n^\ell/\varepsilon)}} + \tfrac{1.1\cdot\sqrt{800}\cdot M(D,d,k,\ell,t,\varepsilon)}{\sqrt{\tau}\sqrt{\log(n^\ell/\varepsilon)}} \cdot \sqrt{\tfrac{|J(v_{i-1})|}{2d}} \cdot 1_{|J(v_{i-1})|>1} \right)^2 \\
& && \text{(due to } \tau = 10^4)\\
&\le \left(\tfrac{M(D,d,k,\ell,t,\varepsilon)}{\sqrt{\log(n^\ell/\varepsilon)}} \right)^2 \sum_{i=1}^{D'} 2 \cdot \left(\frac{4}{\tau D} + \tfrac{968}{\tau} \cdot \tfrac{|J(v_{i-1})|}{2d} \cdot 1_{|J(v_{i-1})|>1} \right) \\
& && \text{(due to } (a+b)^2 \le 2(a^2+b^2))\\
&\le \left(\tfrac{2000\cdot M(D,d,k,\ell,t,\varepsilon)}{\tau\cdot\sqrt{\log(n^\ell/\varepsilon)}} \right)^2 = R^2.
\end{aligned}$$

◁