1 DNFs

The goal of this section is to study the Fourier structure of DNFs and their generalizations, low depth circuits. To recall, a DNF is a boolean formula $f : \{0, 1\}^n \rightarrow \{0, 1\}$ of the form

$$f(x) = C_1(x) \lor \ldots \lor C_m(x)$$

where each $C_i(x)$ is a term, namely an AND of literals (variables or their negations). We denote by $|C_i|$ the number of variables appearing in $C_i$. The width of the DNF is max $|C_i|$, and its size is $\sum |C_i|$. A CNF is very similar to a DNF: it’s the OR of AND of literals. From the Fourier perspective they are equivalent, as you can get a CNF from a DNF by negating it and applying De-Morgan’s formulas. So we will focus on DNFs for concreteness.

DNFs are more powerful than decision trees.

**Claim 1.1.** Let $f : \{0, 1\}^n \rightarrow \{0, 1\}$ be computed by a decision tree of depth $d$ and size $s$. Then it is also computed by a DNF of width $d$ and size $s$.

**Proof.** For each leaf $\ell$ of the decision tree, let $f_\ell(x) = 1$ if $x$ reaches $\ell$ and 0 otherwise. Let $v(\ell) \in \{0, 1\}$ be the value associated to $\ell$. Note that $f_\ell$ is a term whose size is the depth of $\ell$, and that

$$f(x) = \bigvee_{\ell: v(\ell) = 1} f_\ell(x).$$

The total influence of bounded width DNFs is bounded.

**Claim 1.2.** Let $f : \{0, 1\}^n \rightarrow \{0, 1\}$ be computed by a DNF of width $w$. Then $I(f) \leq 2w$.

**Proof.** Recall that $I(f) = \sum I_{f_i}(f)$. Equivalently, it is the total number of sensitive edges $(x, x^{\oplus i})$ where $f(x) \neq f(x^{\oplus i})$, divided by $2^n$. If we restrict to $f(x) = 1$ then it is equivalent to

$$I(f) = 2 \cdot 2^{-n}|\{(x, x^{\oplus i}) : f(x) = 1, f(x^{\oplus i}) = 0\}|.$$

Fix $x$ where $f(x) = 1$. Recall that by assumption $f(x) = C_1(x) \lor \ldots \lor C_m(x)$ where each $C_i$ is a term of width at most $w$. If $f(x) = 1$ then $C_j(x) = 1$ for at least one $j$ (if there are
a few, choose one arbitrarily). If \( f(x^{\oplus i}) = 0 \) then it must be the case that \( x_i \) is one of the variables in \( C_j \). Thus, there are at most \( w \) choices for \( i \) and thus
\[
I(f) \leq 2 \cdot 2^{-n} \cdot 2^n \cdot w = 2w. 
\]
\[\square\]

**Corollary 1.3.** If \( f \) is computed by a DNF of width \( w \) then its spectrum if \( \varepsilon \)-concentrated up to degree \( 2w/\varepsilon \).

As we will see later, the dependence on \( w \) is correct but the dependence on \( \varepsilon \) can be improved.

Usually DNFs of small size, rather than small width, are studied. However we can always approximate the former by the latter.

**Claim 1.4.** Let \( f \) be computed by a DNF of size \( s \). Then for any \( \varepsilon > 0 \), there exists a DNF \( g \) of width \( \log (s/\varepsilon) \) such that \( \Pr[f(x) \neq g(x)] \leq \varepsilon \).

**Proof.** Assume \( f(x) = \bigvee_{i=1}^{m} C_i(x) \), where \( m \leq s \). For \( w = \log (s/\varepsilon) \) let \( g(x) = \bigvee_{i:|C_i| \leq w} C_i(x) \). Then \( f(x) \neq g(x) \) only when \( C_i(x) = 1 \) for some term \( C_i \) of size \( |C_i| > w \). But then
\[
\Pr[f(x) \neq g(x)] \leq \Pr[\exists i, |C_i| > w, C_i(x) = 1] \leq \sum_{i:|C_i| > w} \Pr[C_i(x) = 1] \leq m 2^{-w} \leq s 2^{-w} \leq \varepsilon.
\]
\[\square\]

**Corollary 1.5.** If \( f \) is computed by a DNF of size \( s \) then its spectrum if \( \varepsilon \)-concentrated up to degree \( O(\log (s/\varepsilon)/\varepsilon) \).

**Proof.** Let \( g \) be a DNF of width \( w = \log (s/\varepsilon) \) so that \( \Pr[f \neq g] \leq \varepsilon \). Let \( d = 2w/\varepsilon \). We know that \( \sum_{|S| > d} |\hat{g}(S)|^2 \leq \varepsilon \). We would like to prove a similar bound for \( f \). To that end, let \( h = f - g \). Using the inequality \( (x + y)^2 \leq 2(x^2 + y^2) \) we can bound
\[
\sum_{|S| > d} |\hat{f}(S)|^2 = \sum_{|S| > d} |\hat{g}(S) + \hat{h}(S)|^2 \leq 2 \sum_{|S| > d} |\hat{g}(S)|^2 + 2 \sum_{|S| > d} |\hat{h}(S)|^2.
\]
As we saw, the first sum is bounded by \( 2\varepsilon \). The second sum can be bounded by Parseval as
\[
\sum_{|S| > d} |\hat{h}(S)|^2 \leq \sum_{S} |\hat{h}(S)|^2 = \mathbb{E}[h^2] = 4 \Pr[f(x) \neq g(x)] \leq 4\varepsilon.
\]
Thus in total we get
\[
\sum_{|S| > d} |\hat{f}(S)|^2 \leq 10\varepsilon.
\]
\[\square\]
We will strengthen this result later in two ways: we will improve the dependence on \( \varepsilon \), and show that the total influence of size \( s \) DNFs is \( O(\log s) \).

In light of the Goldreich-Levin algorithm, it would have been useful if poly(\( n \))-size DNFs had their Fourier spectrum concentrated on small sets of size poly(\( n \)) (not necessarily low degree). As we will later see, this is true for sets of size \( n^{O(\log \log n)} \). Mansour conjectured a stronger bound.

**Conjecture 1.6** (Mansour’s conjecture). Let \( f : \{0,1\}^n \to \{0,1\} \) be computed by a DNF of size \( s \). Then there is a collection \( F \) of Fourier coefficients of size \( |F| \leq s^{O(\log 1/\varepsilon)} \) such that \( \sum_{S \notin F} \hat{f}(S)^2 \leq \varepsilon \). A weaker conjecture is: if \( s = \text{poly}(n) \) and \( \varepsilon = O(1) \) then \( |F| = \text{poly}(n) \).

Note that this conjecture implies that poly(\( n \))-size DNFs can be learned from membership queries in time poly(\( n \)), by finding and estimating all their Fourier coefficients \( S \) with \( \hat{f}(S)^2 \geq \varepsilon/|F| \). This turns out to be possible via a different algorithm (Jackson’s algorithm) using totally different techniques.

## 2 Random restrictions

One of the useful tools in studying of the Fourier structure of boolean functions is their behavior under random restrictions. A \( p \)-random restriction of \( f : \{0,1\}^n \to \{0,1\} \) is the boolean function obtained by the following process: for every input variable \( x_i \), with probability \( p \) keep it alive, and otherwise set it to be either 0 or 1 with equal probability (namely \( (1-p)/2 \)). Equivalently, we can first sample the set \( I \) of variables that stayed alive, and then sample uniformly the values \( z \) of the variables outside \( I \). It will be useful to embed the restricted function again in \( \{0,1\}^n \), by setting the variables outside \( I \) to \( z \).

To that end, the following notation is useful. For \( I \subseteq [n] \) and \( z \in \{0,1\}^n \) define \( f_{I,z} : \{0,1\}^n \to \{0,1\} \) as

\[
f_{I,z}(x) = f(y_1, \ldots, y_n),
\]

where

\[
y_i = \begin{cases} x_i & \text{if } i \in I \\ z_i & \text{if } i \notin I \end{cases}.
\]

**Definition 2.1** (Random restriction). The \( p \)-random restriction of \( f : \{0,1\}^n \to \{0,1\} \) is the random variable \( f_{I,z} : \{0,1\}^n \to \{0,1\} \), where \( I \) is sampled by picking \( i \in I \) with probability \( p \) for each \( i \in [n] \) independently, and where \( z \in \{0,1\}^n \) is uniformly chosen.

The Fourier coefficients of \( f \) are tightly related to that of its restriction.

**Lemma 2.2.** Let \( f : \{0,1\}^n \to \{0,1\} \) and let \( f_{I,z} \) be its \( p \)-random restriction. Then for every \( S \subseteq [n] \) it holds that

\[
\mathbb{E}_{I,z}[\hat{f}_{I,z}(S)] = \Pr[S \subseteq I] \cdot \hat{f}(S) = p^{|S|}\hat{f}(S).
\]
and
\[ \mathbb{E}_{I,z}[\hat{f}_{I,z}(S)^2] = \sum_{U \subseteq S \subseteq U} \Pr[U \cap I = S] \cdot \hat{f}(U)^2 = \sum_{U \subseteq S \subseteq U} p^{|S|}(1 - p)^{|U \setminus S|} \cdot \hat{f}(U)^2. \]

**Proof.** For a fixed \( I, z \) the Fourier expansion of \( f_{I,z} \) is
\[ f_{I,z}(x) = \sum_{A \subseteq I, B \subseteq I^c} \hat{f}(A \cup B) x^A z^B. \]
Next, fix \( I \) and randomly sample \( z \). We first compute the average value of \( \hat{f}_{I,z}(S) \):
\[
\begin{align*}
\mathbb{E}_z[\hat{f}_{I,z}(S)] &= \mathbb{E}_{x,z}[f_{I,z}(x) x^S] \\
&= \sum_{A \subseteq I, B \subseteq I^c} \hat{f}(A \cup B) \mathbb{E}_{x,z} [x^A z^B x^S] \\
&= \sum_{A \subseteq I, B \subseteq I^c} \hat{f}(A \cup B) 1_{A=s, B=\emptyset} \\
&= \hat{f}(S) 1_{S \subseteq I}.
\end{align*}
\]
Next, we compute the second moment:
\[
\begin{align*}
\mathbb{E}_z[\hat{f}_{I,z}(S)^2] &= \mathbb{E}_z[(\mathbb{E}_x[f_{I,z}(x) x^S])]^2 \\
&= \mathbb{E}_{z,x,x''}[f_{I,z}(x') f_{I,z}(x'') (x'')^S (x'')^S] \\
&= \sum_{A',A'' \subseteq I, B',B'' \subseteq I^c} \hat{f}(A' \cup B') \hat{f}(A'' \cup B'') \mathbb{E}_{z,x,x''} [(x')^A z^B (x'')^A'' z^B'' (x'')^S (x'')^S] \\
&= \sum_{A',A'' \subseteq I, B',B'' \subseteq I^c} \hat{f}(A' \cup B') \hat{f}(A'' \cup B'') \cdot 1_{A'=A'', A''=S, B''=S} \\
&= \sum_{B \subseteq I^c} \hat{f}(S \cup B)^2 \cdot 1_{S \subseteq I} \\
&= \sum_{U \supseteq S} \hat{f}(U)^2 \cdot 1_{U \cap I = S}.
\end{align*}
\]
The lemma follows by averaging over \( I \) as well. \( \square \)

**Corollary 2.3.** Let \( f : \{0,1\}^n \to \{0,1\} \) and let \( f_{I,z} \) be its \( p \)-random restriction. Then
\[ \mathbb{E}[\text{Inf}_i(f_{I,z})] = p \cdot \text{Inf}_i(f). \]
In particular, \( \mathbb{E}[I(f_{I,z})] = p \cdot I(f). \)

**Proof.** We have
\[
\begin{align*}
\mathbb{E}[\text{Inf}_i(f_{I,z})] &= \sum_{S : i \in S} \mathbb{E}[\hat{f}_{I,z}(S)]^2 \\
&= \sum_{S : i \in S} \sum_{U \subseteq S \subseteq U} p^{|S|}(1 - p)^{|U \setminus S|} \hat{f}(U)^2 \\
&= \sum_{U : i \in U} \hat{f}(U)^2 \left( \sum_{S : i \in S} p^{|S|}(1 - p)^{|U \setminus S|} \right)
\end{align*}
\]
Next observe that
\[
\sum_{S: i \in S \subseteq U} p^{|S|} (1 - p)^{|U \setminus S|} = p \cdot \sum_{S: S \subseteq U \setminus \{i\}} p^{|S|} (1 - p)^{|U \setminus S|} = p (p + (1 - p))^{|U|-1} = p.
\]

Thus
\[
\mathbb{E}[\text{Inf}_i(f_{I,z})] = p \sum_{U: i \in U} \hat{f}(U)^2 = p \cdot \text{Inf}_i(f).
\]

We use this to upper bound the total influence of a DNF, by considering the total influence of its random restrictions.

**Theorem 2.4.** Let \( f : \{0, 1\}^n \to \{0, 1\} \) be computed by a DNF of size \( s \). Then \( I(f) \leq O(\log s) \).

**Proof.** Let \( f(x) = \lor_{i=1}^m C_i(x) \). We will analyze its \((1/2)\)-random restriction. Define the width of \( f_{I,z} \) to be the width of the largest surviving term after the random restriction. If \(|C_i| = w\) then the probability it survives the random restriction (i.e. not become identically zero) is \((3/4)^w\). Thus
\[
\Pr[\text{width}(f_{I,z}) \geq w] \leq \sum_{i: |C_i| \geq w} \Pr[C_i \text{ survives}] \leq m (3/4)^w \leq s(3/4)^w.
\]

if \( \text{width}(f_{I,z}) = w \) then we saw that \( I(f_{I,z}) \leq 2w \). Let \( w_0 = O(\log s) \) so that \( s(3/4)^w_0 \leq 1/2 \). Then
\[
\mathbb{E}[I(f_{I,z})] = \sum_{w \geq 0} \mathbb{E}[I(f_{I,z}) | \text{width}(f_{I,z}) = w] \Pr[\text{width}(f_{I,z}) = w]
\leq \sum_{w \geq 0} 2w \Pr[\text{width}(f_{I,z}) = w]
\leq 2w_0 + \sum_{w > w_0} 2w \cdot s(3/4)^w.
\]

The second term can be bounded by
\[
\sum_{w > w_0} 2w \cdot s(3/4)^w \leq \sum_{w \geq 1} 2(w + w_0) \cdot (1/2)^w \leq \sum_{w \geq 1} 2w (1/2)^w + 2w_0 \sum_{w \geq 1} (1/2)^w \leq O(w_0).
\]

We conclude that
\[
I(f) = 2\mathbb{E}[I(f_{I,z})] \leq O(w_0) = O(\log s).
\]
3 Håstad’s switching lemma

The switching lemma allows for a better analysis for the behaviour of a DNF under random restrictions. Consider first the behaviour of a term of width \( w \) under \( p \)-random restriction. There are three possibilities:

- At least one literal is set to False. This happens with probability \( 1 - \left( \frac{1+p}{2} \right)^w \). In such a case, the entire term is False.
- All literals are set to True. This happens with probability \( \left( \frac{1-p}{2} \right)^w \). In such a case, the entire term is True.
- Some literals are set to True, none to False, and at least one literal remains alive. This happens with probability \( \left( \frac{1+p}{2} \right)^w - \left( \frac{1-p}{2} \right)^w \). In such a case, the term value is not yet determined.

If \( p \ll \frac{1}{w} \) then the third option is very unlikely. The following shows that in such a case, the entire DNF gets fixed with high probability.

**Lemma 3.1** (Baby version of the switching lemma). Let \( f : \{0,1\}^n \rightarrow \{0,1\} \) be computed by a DNF of width \( w \). Let \( f_{I,z} \) be a \( p \)-random restriction of \( f \). Then

\[
\Pr[f_{I,z} \text{ is not a constant function}] \leq O(pw).
\]

The full switching lemma allows to obtain better probability bounds, by considering the decision tree depth of the restriction. For a boolean function \( f \) we denote by \( DT(f) \) the minimal depth of a decision tree computing \( f \).

**Theorem 3.2** (Håstad Switching lemma). Let \( f : \{0,1\}^n \rightarrow \{0,1\} \) be computed by a DNF of depth \( w \). Let \( f_{I,z} \) be a \( p \)-random restriction of \( f \). Then for any \( k \geq 1 \) it holds that

\[
\Pr[DT(f_{I,z}) \geq k] \leq (5pw)^k.
\]

Observe that the baby switching lemma is the \( k = 1 \) version of the full blown switching lemma. We will not prove the switching lemma here. Instead, we will explore corollaries of it on the Fourier structure of DNFs. We start with the following general lemma.

**Lemma 3.3.** Let \( f : \{0,1\}^n \rightarrow \{0,1\} \). Let \( f_{I,z} \) be a \( p \)-random restriction of \( f \), and assume that \( \Pr[DT(f_{I,z}) \geq k] \leq \varepsilon \). Then the spectrum of \( f \) is \( O(\varepsilon) \)-concentrated up to degree \( k/p \).

**Proof.** Let \( d = k/p \). We will bound \( W_{\geq d}(f) = \sum_{S : |S| \geq d} \hat{f}(S)^2 \).

Note that if \( f_{I,z} \) is computed by a decision tree of depth \( < k \), then it has no Fourier coefficients of weight \( k \). Thus

\[
\mathbb{E}[W_{\geq k}(f_{I,z})] \leq \Pr[DT(f_{I,z}) \geq k] \leq \varepsilon.
\]
On the other hand, recall that we computed that
\[ E[\hat{f}_{I,z}(S)^2] = \sum_{U: S \subseteq U} \Pr[U \cap I = S] \cdot \hat{f}(U)^2. \]

Thus if we sum over all \( S \) with \( |S| \geq k \) we get that
\[ E[W_{\geq k}(f_{I,z})] = \sum_U \Pr[|U \cap I| \geq k] \cdot \hat{f}(U)^2. \]

If \( |U| \geq d \) then we can lower bound \( \Pr[|U \cap I| \geq k] \) by
\[ \Pr[|U \cap I| \geq k] = \Pr[\text{Bin}(|U|, p) \geq k] \geq \Pr[\text{Bin}(|U|, p) \geq p|U|] \geq c \]
for some absolute constant \( c > 0 \). Thus
\[ E[W_{\geq k}(f_{I,z})] \geq \sum_{U: |U| \geq d} c \cdot \hat{f}(U)^2 = c \cdot W_{\geq d}(f). \]

Thus we conclude that
\[ W_{\geq d}(f) \leq (1/c) E[W_{\geq k}(f_{I,z})] \leq (1/c) \varepsilon = O(\varepsilon). \]

\[ \square \]

**Corollary 3.4.** Let \( f : \{0, 1\}^n \rightarrow \{0, 1\} \) be computed by a DNF of depth \( w \). Then the spectrum of \( f \) is \( \varepsilon \)-concentrated up to degree \( O(w \log(1/\varepsilon)) \).

**Proof.** Apply the previous lemma with \( p = 1/10w \) and \( k = \log(1/\varepsilon) \). \( \square \)

We next move to study the \( \ell_1 \) norm of DNFs and their low degree truncations. The following claim will be useful.

**Claim 3.5.** Let \( f : \{0, 1\}^n \rightarrow \{0, 1\} \) and let \( f_{I,z} \) be its \( p \)-random restriction. Then
\[ \sum_S p^S |\hat{f}(S)| \leq E \left[ 2^{DT(f_{I,z})} \right]. \]

**Proof.** We will use the fact that a decision tree \( g \) of depth \( k \) had \( \|\hat{g}\|_1 \leq 2^k \). Let \( e(S) = \text{sign}(\hat{f}(S)) \). Then
\[ \sum_S p^S |\hat{f}(S)| = \sum_S e(S)p^S \hat{f}(S) = E \left[ \sum_S e(S)\hat{f}_{I,z}(S) \right] \leq E[\|\hat{f}_{I,z}(S)\|_1] \leq E \left[ 2^{DT(f_{I,z})} \right]. \]
\[ \square \]
Corollary 3.6. Let \( f : \{0, 1\}^n \to \{0, 1\} \) be computed by a DNF of depth \( w \). Then
\[
\sum_{S : |S| \leq k} |\hat{f}(S)| \leq 2(20w)^k.
\]

Proof. Let \( f_{I,z} \) be a \( p \)-restriction of \( f \) for \( p = 1/20w \). Then
\[
\sum_{S} p^{|S|} |\hat{f}(S)| \leq \mathbb{E}[2^{DT(f_{I,z})}] \leq \sum_{k \geq 0} 2^k (5pw)^k \leq 2.
\]
In particular
\[
\sum_{S : |S| \leq k} |\hat{f}(S)| \leq 2(1/p)^k = 2(20w)^k.
\]

This will allow us to almost prove Mansour’s conjecture. We focus on the case of DNFs of width \( O(\log n) \).

Lemma 3.7. Let \( f : \{0, 1\}^n \to \{0, 1\} \) be computed by a DNF of depth \( w \). Then there exists a subset \( \mathcal{F} \) of size \( |\mathcal{F}| \leq n^{O(w log(1/\varepsilon))} \) such that \( \sum_{S \in \mathcal{F}} \hat{f}(S)^2 \leq \varepsilon \). In particular for \( w = O(\log n), \varepsilon = O(1) \) we have \( |\mathcal{F}| = n^{O(\log \log n)} \).

Proof. Let \( k = O(w \log(1/\varepsilon)) \) and let \( g = \sum_{|S| \leq k} \hat{f}(S)x^S \). We have that
\[
\|g\|_1 \leq w^{O(w \log(1/\varepsilon))} \quad \text{and} \quad \|f - g\|_2^2 = \sum_{|S| > k} \hat{f}(S)^2 \leq \varepsilon.
\]
Let \( \tau = \varepsilon/\|g\|_1 \) and take \( \mathcal{F} = \{S : |\hat{f}(S)| \geq \tau\} \). Then \( |\mathcal{F}| \leq (1/\tau)^2 = w^{O(w \log(1/\varepsilon))} \) and
\[
\sum_{S \in \mathcal{F}} \hat{f}(S)^2 \leq \sum_{S : |S| \leq k, |\hat{f}(S)| \leq \tau} |\hat{f}(S)|^2 + \sum_{S : |S| > k} |\hat{f}(S)|^2 \leq \tau \|g\|_1 + \varepsilon \leq 2\varepsilon.
\]

\[\square\]

4 Constant depth circuits

In this section we extend our analysis to constant depth circuits with AND, OR, NOT gates. These extend DNFs and CNFs (which can be viewed as depth two circuits) and give much more expressive power. In general, circuits are DAGs whose leaves (i.e. fan-in zero nodes) correspond to inputs, and whose single root (i.e. fan-out zero node) corresponds to the output, and where internal nodes are labeled by AND, OR gates of unbounded fan-in or NOT gates. We will assume that the circuits are in normal form, as described below. It’s easy to check that any circuit can be transformed to a normal form with at most doubling its depth and increasing it’s size by a constant factor.
Definition 4.1 (Constant depth circuits; normal form). A depth \(d\) circuit in normal form computing a boolean function \(f: \{0, 1\}^n \to \{0, 1\}\) is a layered graph with \(d + 1\) layers as follows:

- The level 0 (input) layer has \(2n\) nodes, corresponding to inputs and their negations.
- There is a single node at level \(d\), corresponding to the output.
- At each intermediate level, all nodes are either AND gates or all OR gates (we may assume these alternate).

The size of the circuit is the number of nodes in it in levels 1, \ldots, \(d - 1\).

We first show how the switching lemma extends from DNFs to low depth circuits. To fix notations, let \(C(s, d)\) denote the class of circuits of size \(s\) and depth \(d\). For technical reasons that will become clear soon, we will also need to assume that the depth 1 nodes have a bounded fan-in. So let \(C(s, d; w)\) denote size \(s\) depth \(d\) depth 1 fan-in \(w\) circuits. Note that as \(C(s, d) \subset C(s + n, d + 1; 1)\), the width restriction to the bottom layer does not really limit the power of the model.

The following lemma shows that a random restriction reduces the depth by one with high probability.

Lemma 4.2. Let \(f \in C(s, d; w)\). Let \(f_{I, z}\) be a \(p\)-random restriction of \(f\) for \(p = 1/10w\). Then

\[
\Pr[f_{I, z} \notin \mathcal{C}(2^k s, d - 1; k)] \leq 2^{-k}s.
\]

Proof. We apply the switching lemma to all nodes of level 2 in the circuit, which compute either DNFs or CNFs. Let \(v\) be such a node, and let \(f^v(x)\) be the DNF computed at the node. We denote by \(f_{I, z}^v\) the random restriction applied to \(f^v\). By the switching lemma we know that

\[
\Pr[\text{DT}(f_{I, z}^v) \geq k] \leq (5pw)^k \leq 2^{-k}.
\]

Thus

\[
\Pr[\text{exists node } v \text{ at level 2 for which DT}(f_{I, z}^v) \geq k] \leq 2^{-k}s.
\]

Assume now that \(\text{DT}(f_{I, z}^v) < k\) for all nodes \(v\) at level 2. Note that a decision tree of depth \(\leq k - 1\) can be expressed as either a DNF or a CNF of width \(\leq k - 1\) and size \(\leq 2^{k-1}\). We may choose either one. If the nodes at level 3 compute AND gates; then we choose CNF; if they compute OR gates; then we choose DNF. In either case we can now merge levels 2 and 3, and hence reduce the depth by one. The bottom fan-in is now \(\leq k\). The size increases to \(\leq 2^ks\). Thus

\[
\Pr[f_{I, z} \notin \mathcal{C}(2^k s, d - 1; k)] \leq 2^{-k}s.
\]

We apply this iteratively to reduce any constant depth circuit to a decision tree.
Theorem 4.3. Let \( f \in \mathcal{C}(s, d; w) \). Fix \( k \geq 1 \) and let \( f_{(I,z)} \) be a \( p \)-random restriction of \( f \) for \( p = \frac{1}{10w} \left( \frac{1}{10k} \right)^{d-2} \). Then
\[
\Pr[DT(f_{I,z}) \geq k] \leq 2^{-k}s.
\]

Proof. The theorem follows by applying random restrictions iteratively to reduce the depth, and noting that applying a \( p \) and then a \( p' \) random restriction iteratively is the same as simply applying a \( p \cdot p' \) random restriction. So, let \( f_0 = f \) and let \( f_i \) be a \( p_i \)-random restriction of \( f_{i-1} \). We set \( p_0 = (1/10w) \) and \( p_i = (1/10k) \) for \( i \geq 1 \), where \( k \) will be chosen later. So define the events \( E_1, \ldots, E_{d-2} \) as
\[
E_i = [f_i \in \mathcal{C}(s_i, d - i; w)]
\]
and \( E_0 = [f_i \in \mathcal{C}(s, d - i; w)] \).

In fact, the sizes \( s_i \) don’t really matter for the analysis. We need a slightly more refined analysis of the previous lemma. Note that the bound really depends only on the number of nodes in level 2, and not the total number of nodes. Moreover, whenever the depth reduction is successful, the number of nodes in level \( i \) for \( i \geq 2 \) in the restriction \( f_{i-1} \) equals the number of nodes of level \( i + 1 \) in \( f \). Thus, if we let \( m_i \) denote the number of nodes in level \( i \) in the circuit computing \( f \), then we in fact have
\[
\Pr[\neg E_i | E_{i-1}] \leq m_{i+1}2^{-k}.
\]
Finally, define the event
\[
E_{d-1} = [DT(f_{d-1}) \geq k].
\]
Then also
\[
\Pr[\neg E_{d-1} | E_{d-2}] \leq 2^{-k}.
\]
Thus we can conclude that
\[
\Pr[DT(f_{d-1}) \geq k] \leq \sum m_i2^{-k} \leq s2^{-k}.
\]
The final random restriction is a \( p \)-random restriction for \( p = \frac{1}{10w} \left( \frac{1}{10k} \right)^{d-2} \).

Setting \( k = \log(s/\varepsilon) \) in the previous theorem, and recalling that \( \mathcal{C}(s, d) \subset \mathcal{C}(s, d + 1; 1) \) we obtain the following cleaner form.

Theorem 4.4. Let \( f \in \mathcal{C}(s, d) \). For \( \varepsilon > 0 \) let \( f_{(I,z)} \) be a \( p \)-random restriction of \( f \) for \( p = \frac{1}{10} \left( \frac{1}{10 \log(s/\varepsilon)} \right)^{d-1} \). Then
\[
\Pr[DT(f_{I,z}) \geq \log(1/\varepsilon)] \leq \varepsilon.
\]

Linial, Mansour and Nisan (LMN) proved a remarkable theorem: the Fourier spectrum of small depth circuits is concentrated in the first few levels.

Theorem 4.5 (LMN). Let \( f : \{0,1\}^n \to \{0,1\} \) with \( f \in \mathcal{C}(s, d) \). Then the spectrum of \( f \) is \( \varepsilon \)-concentrated up to degree \( O(\log(s/\varepsilon))^{d-1} \log(1/\varepsilon) \).
Proof. Recall that we proved a general lemma stating that if $f_{I,z}$ is a $p$-random restriction of $f$ with $\Pr[\text{DT}(f_{I,z}) \geq k] \leq \varepsilon$ then the spectrum of $f$ is $O(\varepsilon)$-concentrated up to degree $k/p$. Apply it with the corollary we just proved.

One can deduce a similar bound on the total sensitivity. We leave the proof as homework.

**Theorem 4.6.** Let $f : \{0, 1\}^n \rightarrow \{0, 1\}$ with $f \in C(s, d)$. Then $I(f) \leq O(\log s)^{d-1}$.

We give two corollaries of the LMN theorem. The first that polynomial size small depth circuits can be learned from uniform samples in quasi-polynomial time.

**Corollary 4.7.** Let $f : \{0, 1\}^n \rightarrow \{0, 1\}$ be computed by a circuit of size $\text{poly}(n)$ and depth $d$. Then $f$ can be $\varepsilon$-learned from random samples for $\varepsilon = 1/\text{poly}(n)$ in time $\exp((\log n)^{d+1})$.

Proof. Apply LMN and the Fourier sampling algorithm.

The next corollary shows that small low depth circuits cannot approximate the parity function $\text{PARITY}(x) = x_1 \oplus \ldots \oplus x_n$.

**Corollary 4.8.** Let $f : \{0, 1\}^n \rightarrow \{0, 1\}$ be computed by a circuit of size $s$ and depth $d$. Assume that $\Pr[f = \text{PARITY}] \geq 2/3$. Then $s \geq 2^{O(n^{1/d-1})}$.

Proof. If $\Pr[f = \text{PARITY}] \geq 1/2 + \varepsilon$ then $\hat{f}[n] = 2\varepsilon$. So it cannot be $\varepsilon^2$ concentrated up to level $n - 1$. 