An entropy proof of the switching lemma and tight bounds on the decision-tree size of AC^0

Benjamin Rossman*
University of Toronto

November 3, 2017

Abstract

We first give a simple entropy argument showing that every m-clause DNF with expected value $\lambda \in [0,1]$ under the uniform distribution has average sensitivity (a.k.a. total influence) at most $2\lambda \log(m/\lambda)$. Using a similar idea, we then show the following switching lemma for an m-clause DNF (or CNF) formula F:

$$\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(F\restriction \mathbf{R}_p) \geq t\ ] \leq O(p\log(m+1))^t. \tag{1}$$

for all $p \in [0,1]$ and $t \in \mathbb{N}$ where $\mathbf{R}_p$ is the p-random restriction and $\mathsf{DT}_{\mathsf{depth}}(\cdot)$ denotes decision-tree depth. Our proof replaces the counting arguments in previous proofs of Håstad's $O(pw)^t$ switching lemma for width-w DNFs [5, 9, 2] with an entropy argument that naturally applies to unbounded-width DNFs with a bounded number of clauses. With respect to AC^0 circuits, our m-clause switching lemma has similar applications as Håstad's width-w switching lemma, including a $2^{\Omega(n^{1/(d-1)})}$ lower bound for PARITY.

An additional result of this paper extends inequality (1) to AC^0 circuits via a combination of Håstad's switching and multi-switching lemmas [5, 6]. For boolean functions $f : \{0,1\}^n \to \{0,1\}$ computable by AC^0 circuits of depth d and size s, we show that

$$\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(f\restriction \mathbf{R}_p) \geq t\ ] \leq (p \cdot O(\log s)^{d-1})^t \tag{2}$$

for all $p \in [0,1]$ and $t \in \mathbb{N}$. As a corollary, we obtain a tight bound on decision-tree size

$$\mathsf{DT}_{\mathsf{size}}(f\restriction \mathbf{R}_p) \leq O(2^{(1-\varepsilon)n}) \quad \text{where } \varepsilon = 1/O(\log s)^{d-1}. \tag{3}$$

Qualitatively, (2) strengthens a similar inequality of Tal [12] with degree in place of $\mathsf{DT}_{\mathsf{depth}}$, and (3) strengthens a similar inequality of Impagliazzo, Matthews and Paturi [7] with $\mathsf{subcube\ partition\ number}$ in place of $\mathsf{DT}_{\mathsf{size}}$.

*Supported by NSERC and a Sloan Research Fellowship

1 Introduction

Håstad's switching lemma [5] is a cornerstone of circuit complexity. Recall that a *DNF formula* is a disjunction $F = C_1 \vee \cdots \vee C_m$ where each clause C_ℓ is a conjunction of literals (variables x_i or their negations $\neg x_i$). The *width* of F is the maximum number of literals in any clause C_ℓ. The switching lemma gives an exponential tail bound on the decision-tree depth of the function $F\restriction \mathbf{R}_p$ (i.e., F under the p-random restriction $\mathbf{R}_p$) when $p \leq 1/O(\text{width}(F))$.

Theorem 1 (Håstad's Switching Lemma [5])**.** *If F is a width-w DNF formula, then*

$$\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(F\restriction \mathbf{R}_p) \geq t\ ] = O(pw)^t$$

for all $p \in [0,1]$ and $t \in \mathbb{N}$.

The first result of this paper is a switching lemma for m-clause DNFs.

Theorem 2 (Switching Lemma for m-Clause DNFs)**.** *If F is an m-clause DNF formula, then*

$$\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(F\restriction \mathbf{R}_p) \geq t\ ] = O(p \log(m+1))^t$$

for all $p \in [0,1]$ and $t \in \mathbb{N}$. (For $t \geq m^{1-\Omega(1)}$, we obtain a slightly stronger bound $O(p \log(\frac{m}{t}+2))^t$.)

Theorems 1 and 2 are closely related, though incomparable.[1] The two switching lemmas have similar applications with respect to AC^0 circuits, including a $2^{O(n^{1/(d-1)})}$ lower bound for PARITY. However, more than the result itself, Theorem 2 is interesting for its proof technique, which replaces the counting arguments in previous proofs of Theorem 1 [5, 9, 2] with a novel entropy argument (see the discussion in Remark 10). This new proof technique directly generalizes to a certain class of *p-pseudorandom restrictions* where the previous counting arguments seem to break down (see the discussion in §4.1).

The second result of this paper extends Theorem 2 to higher-depth circuits and slightly sharpens the previous knowledge of AC^0.

Theorem 3 (Criticality and Decision-Tree Size of AC^0 Circuits)**.** *If $f : \{0,1\}^n \to \{0,1\}$ is computable by an AC^0 circuit of depth d and size s, then setting $r = 1/O(\log s)^{d-1}$ we have*

(i) $\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(f\restriction \mathbf{R}_p) \geq t\ ] \leq (pr)^t$ *for all $p \in [0,1]$ and $t \in \mathbb{N}$,*

(ii) $\mathsf{DT}_{\mathsf{size}}(f) = O(2^{(1-\frac{1}{r})n})$.

Previously, Tal [12] had shown that (i) holds with degree in place of $\mathsf{DT}_{\mathsf{depth}}$, and Impagliazzo, Matthews and Paturi [7] had shown that (ii) holds with $\mathsf{subcube\ partition\ number}$ in place of $\mathsf{DT}_{\mathsf{size}}$.[2] Theorem 3 is ultimately proved by a (mostly straightforward) application of Håstad's switching and multi-switching lemmas [5, 6]. Of independent interest, we introduce a notion of *criticality* of a boolean function f (the threshold value of p below which $\mathsf{DT}_{\mathsf{depth}}(f\restriction \mathbf{R}_p)$ has an exponential tail bound) and observe a connection to decision-tree size.

[1] In the case of $t = O(\log(m+1))$, Theorem 2 can be derived from Theorem 1 by truncating clauses to width $\log(m+1)$; in this case, the $m^{-O(1)}$ approximation error will be less than $\exp(-t)$. However, this reduction fails for $t \gg \log(m+1)$; this is significant in the context of *criticality* (see §5). In the other direction, Theorem 1 reduces to Theorem 2 in the (typical) special case where F is a disjunction or conjunction of $2^{O(w)}$ many depth-w decision trees (see Corollary 14).

[2] Note that $\mathsf{degree} \leq \mathsf{DT}_{\mathsf{depth}}$ and $\mathsf{subcube\ partition\ number} \leq \mathsf{DT}_{\mathsf{size}}$. The main objective of [7] is a satisfiability algorithm for AC^0. The bound on $\mathsf{subcube\ partition\ number}$ obtained along the way might in fact arise from a decision tree; however, this is difficult to ascertain. Their bound on $\mathsf{subcube\ partition\ number}$ is actually $O(2^{(1-\frac{1}{r'})n})$ where $r' = 1/O(\log(s/n))^{d-1}$; quantitatively, this is better than $O(2^{(1-\frac{1}{r})n})$ for almost-linear size $s \leq n^{1+o(1)}$.

Overview. In §2 we state some preliminary definitions. In §3, as a warm-up to Theorem 2, we present a simple entropy proof that m-clause DNFs have average sensitivity at most $2m$. We then prove Theorem 2 in §4. In §5 we introduce the notion of *criticality* and describe its connection to Theorem 3. (Proofs of the results of this section are included in appendices.) We conclude in §7 by mentioning some open questions raised by this work.

2 Preliminary Definitions

Let $\mathbb{N} := \{0, 1, 2, \dots\}$ and $\mathbb{N}_+ := \{1, 2, \dots\}$. For $s \in \mathbb{N}$, let $[s] := \{1, \dots, s\}$. For a set S and $t \in \mathbb{N}$, $\binom{S}{t}$ is the set of t-element subsets of S. $\ln(\cdot)$ and $\log(\cdot)$ are logarithm with base e and 2, respectively. The *entropy* of a distribution $\mu = (\mu_1, \dots, \mu_m)$ with $\mu_i \geq 0$ and $\sum_{i \in [m]} \mu_i = 1$ is the quantity $\mathbb{H}(\mu) := \sum_{i \in [m]} \mu_i \log(1/\mu_i)$, which is always at most $\log(m)$.

Throughout this paper, we fix an arbitrary positive integer n and regard $[n]$ as the set of variable indices for elements of the hypercube $\{0,1\}^n$. A *boolean function* is a function $f : \{0,1\}^n \to \{0,1\}$.

A *restriction* is a partial assignment $\sigma \in \{0,1\}^S$ where $S \subseteq [n]$. We write $\mathrm{Dom}(\sigma) := S$ and $\mathrm{Stars}(\sigma) := [n] \setminus S$. For restrictions $\sigma \in \{0,1\}^S$ and $\tau \in \{0,1\}^T$ with disjoint supports $S \cap T = \emptyset$, we write $\sigma \cup \tau$ for the combined restriction in $\{0,1\}^{S \cup T}$. For $p \in [0,1]$, the *p-random restriction*, denoted $\mathbf{R}_p$, is a uniform random element $\{0,1\}^{\boldsymbol{I}}$ where $\boldsymbol{I}$ is a $1-p$-binomial random subset $[n]$ (which includes each $i \in [n]$ independently with probability $1-p$).

A *decision tree* is a rooted binary tree whose internal nodes (i.e., non-leaves) are labeled by variables and whose leaves are labeled by output values (by default, either 0 or 1). The *depth* of a decision tree is the maximum number of variables queried on a branch. The *size* of a decision tree is the number of leaves. For a boolean function f, we denoted by $\mathsf{DT}_{\mathsf{depth}}(f)$ and $\mathsf{DT}_{\mathsf{size}}(f)$ the minimum depth and size of a decision tree that computes f.

In this paper, *circuits* refers to single-output, alternating AC^0 circuits; by default, we assume that inputs to circuits are labeled by literals (variables x_i or negated variables $\neg x_i$). The *depth* of a circuit is the maximum number of AND and OR gates on any input-to-output path. The *size* of a circuit is the number of gates. Under this definition, depth-0 circuits have size 0 and depth-1 circuits have size 1.

A *formula* is a circuit with the structure of a tree. The special case of depth-2 formulas are known as *DNFs* (OR $\circ$ AND formulas) and *CNFs* (AND $\circ$ OR formulas). Formally, a DNF formula is an ordered sequence of clauses written in the form $F = C_1 \vee \cdots \vee C_m$ where each C_ℓ is a conjunction of literals. The *width* of a DNF is the maximum number of variables in a clause C_ℓ.

3 Warm-Up: Average Sensitivity

As a warm-up, we present a simple proof that every m-clause DNF F with expected value $\lambda \in [0,1]$ has average sensitivity at most $\max\{2\log(m+1), 2\lambda\log(m/\lambda)\}$. Up to an $1 + o(1)$ factor, these bounds can be derived from known results on the average sensitivity of width-w DNFs (see Remark 5). However, our proof involves different argument based on the entropy of the "first witness function" associated with F. This argument was the starting point for our alternative proof of the switching lemma and provides a simple illustration of the underlying principle.

Recall the definitions of *sensitivity* and *average sensitivity*. For a function f with domain $\{0,1\}^n$

and a point $x \in \{0,1\}^n$, let

$$\mathsf{S}(f,x) := |\{i \in [n] : f(x) \neq f(x \oplus i)\}| \quad \text{and} \quad \mathsf{AS}(f) := \mathop{\mathbb{E}}_{x\in\{0,1\}^n}[\ \mathsf{S}(f,x)\ ].$$

The *expected value* of f is $\mathbb{E}_{x\in\{0,1\}^n}[\ f(x)\ ]$.

Theorem 4. *Every m-clause DNF with expected value λ has average sensitivity at most* $\min\{2\log(m+1), 2\lambda\log(m/\lambda)\}$.

Proof. Let $F = C_1 \vee \cdots \vee C_m$ be an m-clause DNF. Let $\widetilde{F} : \{0,1\}^n \to [m+1]$ be the "first witness function" mapping $x \in \{0,1\}^n$ to the index of the first satisfied clause if any, and otherwise to $m+1$. Let

$$\mathsf{S}_<(\widetilde{F},x) := |\{i \in [n] : \widetilde{F}(x) < \widetilde{F}(x \oplus i)\}| \quad \text{and} \quad \mathsf{AS}_<(\widetilde{F}) := \mathop{\mathbb{E}}_{x\in\{0,1\}^n}[\ \mathsf{S}_<(\widetilde{F},x)\ ].$$

Observe that $\mathsf{AS}(F) \le \mathsf{AS}(\widetilde{F}) = 2{\cdot}\mathsf{AS}_<(\widetilde{F})$.

Let $\mu = (\mu_1, \ldots, \mu_{m+1})$ be the probability distribution induced by $\widetilde{F}$ under the uniform distribution on $\{0,1\}^n$, that is, $\mu_\ell := \mathbb{P}_{x\in\{0,1\}^n}[\ \widetilde{F}(x) = \ell\ ]$. For each $\ell \in [m]$, we have

$$\begin{aligned}
2^{\mathbb{E}_{y\in\widetilde{F}^{-1}(\ell)}[\ \mathsf{S}_<(\widetilde{F},y)\ ]} &\le \mathop{\mathbb{E}}_{y\in\widetilde{F}^{-1}(\ell)}[\ 2^{\mathsf{S}_<(\widetilde{F},y)}\ ] && \text{by Jensen's inequality}\\
&\le 2^{|C_\ell|} && \text{since } \mathsf{S}_<(\widetilde{F},y) \le |C_\ell| \text{ for all } y \in \widetilde{F}^{-1}(\ell)\\
&\le \frac{1}{\mu_\ell} && \text{since } \mu_\ell \le \mathop{\mathbb{P}}_{x\in\{0,1\}^n}[\ C_\ell(x) = 1\ ] = 2^{-|C_\ell|}.
\end{aligned}$$

Therefore, $\mathbb{E}_{y\in\widetilde{F}^{-1}(\ell)}[\ \mathsf{S}_<(\widetilde{F},y)\ ] \le \log(1/\mu_\ell)$.

Using the fact that μ has entropy at most $\log(m+1)$, we have

$$\begin{aligned}
\mathsf{AS}_<(\widetilde{F}) &= \mathop{\mathbb{E}}_{x\in\{0,1\}^n}[\ \mathsf{S}_<(\widetilde{F},x)\ ]\\
&= \sum_{\ell\in[m]} \mu_\ell \mathop{\mathbb{E}}_{y\in\widetilde{F}^{-1}(\ell)}[\ \mathsf{S}_<(\widetilde{F},y)\ ]\\
&\le \sum_{\ell\in[m]} \mu_\ell \log(1/\mu_\ell) \ \le\ \sum_{\ell\in[m+1]} \mu_\ell \log(1/\mu_\ell) \ =\ \mathbb{H}(\mu) \ \le\ \log(m+1).
\end{aligned}$$

We conclude that $\mathsf{AS}(F) \le 2\log(m+1)$.

If F has expected value λ, then letting $\mu'_\ell := \mu_\ell/\lambda$ (and noting that $\lambda = \sum_{\ell\in[m]} \mu_\ell$), we have

$$\sum_{\ell\in[m]} \mu_\ell \log(1/\mu_\ell) = \lambda \sum_{\ell\in[m]} \mu'_\ell\Big(\log(1/\mu'_\ell) - \log(\lambda)\Big) = \lambda\Big(\mathbb{H}(\mu') - \log(\lambda)\Big) \le \lambda\log(m/\lambda).$$

This gives the bound $\mathsf{AS}(F) \le 2\lambda\log(m/\lambda)$. □

For $k, t \in \mathbb{N}$, observe that the function $\mathrm{PARITY}(x_1,\ldots,x_k) \wedge \mathrm{AND}(x_{k+1},\ldots,x_{k+t})$ is equivalent to a DNF with $m := 2^k$ clauses and has expected value $\lambda := (1/2)^{t+1}$ and average sensitivity $2\lambda(\log(m/\lambda) - 1)$ $(= 2\lambda(k+t))$. This shows that Theorem 4 is essentially tight for $\lambda \in [0, \frac{1}{2}]$.

Remark 5. The average sensitivity of a width-w DNF with expected value λ is known to be at most the minimum of w (Amano [1]), $2\lambda w$ (Boppana [4]) and $2(1-\lambda)w/\log(1/(1-\lambda))$ (Traxler [13]). Each of these bounds is tight for a certain values of λ. Extending all three bounds, Scheder and Tan [11] proved an upper bound of $\beta(\lambda)w$ for a certain piecewise linear function $\beta : [0,1] \to [0,1]$; this bound is asymptotically tight for all values of λ. By approximating any m-clause by a DNF of width $\lceil \log m \rceil$, they also observe that $(1+o(1))\beta(\lambda)\log(m+1)$ is an upper bound on the average sensitivity of m-clause DNFs.

Remark 6. A weak converse to Theorem 4: Keller and Lifshitz [8] recently showed that every boolean function with expected value λ and average sensitivity at most $2\lambda\log(m/\lambda)$ is $\varepsilon\lambda$-approximated by a DNF of size $2^{m^{O(1/\varepsilon)}}$.

4 Switching Lemma for m-Clause DNFs

The next lemma is a generalization of the fact that the Shannon entropy of a probability distribution μ is at most $\log|\mathrm{Supp}(\mu)|$. Lemma 7 involves the entropy-like quantity $\sum_i \mu_i \log(1/\mu_i)^t$ where $t \in \mathbb{N}$, of which Shannon entropy is the case $t = 1$.

Lemma 7. *For all $s, t \in \mathbb{N}_+$ and $\mu_1, \ldots, \mu_s \in [0,1]$,*

$$\sum_{i=1}^{s} \mu_i \left(\frac{\ln(1/\mu_i)}{t}\right)^t \le \left(\frac{\ln(s)}{t}\right)^t + 2.$$

Proof. The function $x(\ln(1/x)/t)^t$ has its maximum value e^{-t} at $x = e^{-t}$. If $s < 2e^t$, then

$$\sum_{i=1}^{s} \mu_i \left(\frac{\ln(1/\mu_i)}{t}\right)^t \le se^{-t} < 2.$$

So we may assume that $s \ge 2e^t$. Let

$$\begin{aligned} r &:= |\{i \in [s] \,:\, \mu_i \ge e^{-t}\}| && \le e^t, \\ \eta &:= \mathop{\mathbb{E}}_{i\in[s]\,:\,\mu_i<e^{-t}}[\ \mu_i\ ] && \le 1/(s-r) \le 1/(s-e^t) \le e^{-t}. \end{aligned}$$

Since $x(\ln(1/x)/t)^t$ is concave and increasing in the interval $[0, e^{-t}]$, by Jensen's inequality

$$\mathop{\mathbb{E}}_{i\in[s]\,:\,\mu_i<e^{-t}} \left[\ \mu_i \left(\frac{\ln(1/\mu_i)}{t}\right)^t\ \right] \le \eta \left(\frac{\ln(1/\eta)}{t}\right)^t \le \frac{1}{s-r}\left(\frac{\ln(s-r)}{t}\right)^t.$$

Therefore,

$$\begin{aligned} \sum_{i=1}^{s} \mu_i \left(\frac{\ln(1/\mu_i)}{t}\right)^t &\le \sum_{i\in[s]\,:\,\mu_i<e^{-t}} \mu_i \left(\frac{\ln(1/\mu_i)}{t}\right)^t + \sum_{i\in[s]\,:\,\mu_i\ge e^{-t}} \mu_i \left(\frac{\ln(1/\mu_i)}{t}\right)^t \\ &\le (s-r) \mathop{\mathbb{E}}_{i\in[s]\,:\,\mu_i<e^{-t}} \left[\ \mu_i \left(\frac{\ln(1/\mu_i)}{t}\right)^t\ \right] + re^{-t} \\ &\le \left(\frac{\ln(s)}{t}\right)^t + 1. \end{aligned}$$

□

For the rest of this section, we fix an m-clause DNF formula $F = C_1 \vee \cdots \vee C_m$. We also fix arbitrary $p \in [0,1]$ and $t \in \mathbb{N}_+$. For $\ell \in [m]$, let $V_\ell \subseteq [n]$ be the set of variables on which C_ℓ depends (i.e., C_ℓ is a conjunction of literals over V_ℓ). For uniform random $\boldsymbol{x} \in \{0,1\}^n$, note that $\mathbb{P}[\ C_\ell(\boldsymbol{x}) = 1\ ] = 2^{-|V_\ell|}$. For $\ell_1, \ldots, \ell_k \in [m]$, note that $\mathbb{P}[\ C_{\ell_1}(\boldsymbol{x}) = \cdots = C_{\ell_k}(\boldsymbol{x}) = 1\ ]$ is either 0 or $2^{-|V_{\ell_1} \cup \cdots \cup V_{\ell_m}|}$ according to whether or not $C_{\ell_1} \wedge \cdots \wedge C_{\ell_k}$ is satisfiable.

As a matter of notation, for $\ell \in [m]$ and a restriction ϱ, let

$$C_\ell(\varrho) := \begin{cases} 0 & \text{if } C_\ell{\restriction}\varrho \equiv 0, \\ 1 & \text{if } C_\ell{\restriction}\varrho \equiv 1, \\ * & \text{otherwise (if } C_\ell{\restriction}\varrho \text{ is nonconstant).} \end{cases}$$

Similar to all known proofs of Håstad's switching lemma for width-w DNFs, our proof of Theorem 2 analyzes the canonical decision tree for $F{\restriction}\varrho$, defined below.

Definition 8. The *canonical decision tree* of $F{\restriction}\varrho$, denoted $\mathrm{CDT}(F{\restriction}\varrho)$, is defined inductively as follows:

- If $C_1(\varrho) = \cdots = C_m(\varrho) = 0$ or there exists $\ell \in [m]$ such that $C_1(\varrho) = \cdots = C_{\ell-1}(\varrho) = 0$ and $C_\ell(\varrho) = 1$, then output 0 or 1 accordingly.
- Otherwise, let $\ell \in [m]$ be unique index such that $C_1(\varrho) = \cdots = C_{\ell-1}(\varrho) = 0$ and $C_\ell(\varrho) = *$. Let $I := V_\ell \setminus \mathrm{Dom}(\varrho)$ be the set of variables on which $C_\ell{\restriction}\varrho$ depends. (Note that I is nonempty.) Query all variables in I, receiving answers $\sigma \in \{0,1\}^I$. Proceed as the canonical decision tree of $F{\restriction}\varrho\sigma$.

Definition 9. For $k \in \mathbb{N}_+$ and $\vec{t} = (t_1, \ldots, t_k) \in \mathbb{N}_+^k$, we say that a restriction ϱ is *$\vec{t}$-bad* with respect to F if there exists a sequence $\vec{\ell} = (\ell_1, \ldots, \ell_k)$ where $1 \le \ell_1 < \cdots < \ell_k \le m$ such that there exists a branch in $\mathrm{CDT}(F{\restriction}\varrho)$ which first queries t_1 variables from C_{ℓ_1}, then queries t_2 variables from C_{ℓ_2}, and so on up to querying t_k variables from C_{ℓ_k}.

In addition to the sequence $\vec{\ell} = (\ell_1, \ldots, \ell_k)$ of clause indices, such a $\vec{t}$-bad branch in $\mathrm{CDT}(F{\restriction}\varrho)$ is associated with data

$$\vec{I} = (I_1, \ldots, I_k), \quad \vec{\sigma} = (\sigma_1, \ldots, \sigma_k), \quad \vec{\tau} = (\tau_1, \ldots, \tau_k)$$

where

- $I_j = V_{\ell_j} \setminus (\mathrm{Dom}(\varrho) \cup I_{\ell_1} \cup \cdots \cup I_{\ell_{j-1}})$ is the set of variables queried from clause C_{ℓ_j},
- $\sigma_j \in \{0,1\}^{I_j}$ is the restriction consisting of answers to queries I_j in the given $\vec{t}$-bad branch,
- $\tau_j \in \{0,1\}^{I_j}$ is unique restriction such that $C_{\ell_j}(\varrho\sigma_1 \ldots \sigma_{j-1}\tau_j) = 1$ (i.e., τ_j is the subclause of C_{ℓ_j} over variables I_j).

Observe that data $\vec{\ell}, \vec{I}, \vec{\sigma}, \vec{\tau}$ satisfy the following three properties:

(i) $I_1 \cup \cdots \cup I_k \subseteq \mathrm{Stars}(\varrho)$,

(ii) $I_j \in \binom{V_{\ell_j} \setminus (V_{\ell_1} \cup \cdots \cup V_{\ell_{j-1}})}{t_j}$ for all $j \in [k]$,

(iii) $C_1(\varrho) = \cdots = C_{\ell_1-1}(\varrho) = 0$ and $C_{\ell_1}(\varrho\tau_1) = 1$
$C_{\ell_1+1}(\varrho\sigma_1) = \cdots = C_{\ell_2-1}(\varrho\sigma_1) = 0$ and $C_{\ell_2}(\varrho\sigma_1\tau_2) = 1$ and
$\vdots$
$C_{\ell_{k-1}+1}(\varrho\sigma_1\ldots\sigma_{k-1}) = \cdots = C_{\ell_k-1}(\varrho\sigma_1\ldots\sigma_{k-1}) = 0$ and $C_{\ell_k}(\varrho\sigma_1\ldots\sigma_{k-1}\tau_k) = 1$.

Remark 10 (**Overview and comparison with Håstad's switching lemma**)**.** The next two paragraphs provide a high-level overview of the proof of Theorem 2 and a comparison with previous proofs of Håstad's switching lemma (Theorem 1). Nothing essential is lost in skipping directly to Lemma 11.

In the setting where F is a width-w DNF with arbitrarily many clauses, Razborov's proof of Håstad's switching lemma [9] is based on an analysis of the function that maps each $\vec{t}$-bad restriction ϱ to the extended restriction $\varrho\tau_1\ldots\tau_k$. This function $\varrho \mapsto \varrho\tau_1\ldots\tau_k$ is shown to be $O(w)^t$-to-1 (by cleverly constructing a second function $\varrho \mapsto \mathsf{Code}(\varrho)$ with the property that $\varrho \mapsto (\varrho\tau_1\ldots\tau_k, \mathsf{Code}(\varrho))$ is 1-to-1 and $|\mathrm{Range}(\mathsf{Code})| = O(w)^t$). Use the fact that $\mathbb{P}[\ \mathbf{R}_p = \varrho\ ] = (\frac{2p}{1-p})^t\,\mathbb{P}[\ \mathbf{R}_p = \varrho\tau_1\ldots\tau_k\ ]$, it directly follows that $\mathbb{P}[\ \mathbf{R}_p$ is $\vec{t}$-bad $\ ] = O(pw)^t$. The bound $\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(F\restriction\mathbf{R}_p) \geq t\ ] = O(pw)^t$ of Theorem 1 then follows from a union bound over $O(1)^t$ choices of $k \in \mathbb{N}_+$ and $\vec{t} \in \mathbb{N}_+^k$ with $t_1 + \cdots + t_k = t$.

In the present setting where F has m clauses of unbounded width, we also essentially consider the map $\varrho \mapsto \varrho\tau_1\ldots\tau_k$ over $\vec{t}$-bad restriction ϱ. However, in lieu of the previous counting argument which bounds the size of preimages of this map, our proof of Theorem 2 involves an entropy argument. We consider a family of probability distributions μ, each supported on increasing sequences $\vec{\ell} = (\ell_1, \ldots, \ell_k)$ of clause indices in $[m]$. (Roughly speaking, each distribution μ in this family corresponds to a Razborov-style decoding procedure applied to a uniform random element $\boldsymbol{x} \in \{0,1\}^n$, where $\mu(\vec{\ell})$ is the probability that the decoding procedure visits clauses $C_{\ell_1}, \ldots, C_{\ell_k}$.) After some manipulations, we end up with a bound $\mathbb{P}[\ \boldsymbol{\varrho}$ is $\vec{t}$-bad $\ ] \leq O(p)^t \cdot \max_\mu \sum_{\vec{\ell}} \mu(\vec{\ell})\big(\frac{1}{t}\log(1/\mu(\vec{\ell}))\big)^t$. Our final bound $O(p\log(m+1))^t$ then follows from the entropy-like inequality Lemma 7, together with the fact that $|\mathrm{Supp}(\mu)| \leq \binom{m}{k} \leq m^t$ for each distribution μ.

Lemma 11 (Main Lemma)**.** *For all $k \in \mathbb{N}_+$ and $\vec{t} = (t_1, \ldots, t_k) \in \mathbb{N}_+^k$ with $t = t_1 + \cdots + t_k$,*

$$\mathbb{P}[\ \mathbf{R}_p \textit{ is } \vec{t}\textit{-bad w.r.t. } F\ ] \leq (4ep\log(e^2 m))^t = O(p\log(m+1))^t.$$

Proof. For better readability, we write $\boldsymbol{\varrho}$ instead of $\mathbf{R}_p$ for the p-random restriction. Taking a union bound over possible choices of data $\vec{\ell}, \vec{I}, \vec{\sigma}, \vec{\tau}$ and exploiting properties (i)–(iii) of Definition 9, we have

$$\mathbb{P}[\ \boldsymbol{\varrho} \text{ is } \vec{t}\text{-bad } \ ] \leq \sum_{\ell_1=1}^{m-k+1} \sum_{\substack{I_1 \in \binom{V_{\ell_1}}{t_1} \\ \sigma_1,\tau_1 \in \{0,1\}^{I_1}}} \sum_{\ell_2=\ell_1+1}^{m-k+2} \sum_{\substack{I_2 \in \binom{V_{\ell_2} \setminus V_{\ell_1}}{t_2} \\ \sigma_2,\tau_2 \in \{0,1\}^{I_2}}} \cdots \sum_{\ell_k=\ell_{k-1}+1}^{m} \sum_{\substack{I_k \in \binom{V_{\ell_k} \setminus (V_{\ell_1} \cup \cdots \cup V_{\ell_{k-1}})}{t_k} \\ \sigma_k,\tau_k \in \{0,1\}^{I_k}}} \beta_{\vec{\sigma}}(\vec{\ell})$$

$$\leq \sum_{\ell_1} \max_{\sigma_1} \sum_{\ell_2} \max_{\sigma_2} \cdots \sum_{\ell_k} \max_{\sigma_k} \alpha(\vec{\ell})\beta_{\vec{\sigma}}(\vec{\ell}) \tag{4}$$

where

$$\alpha(\vec{\ell}) := 2^t \binom{|V_{\ell_1}|}{t_1} \binom{|V_{\ell_2} \setminus V_{\ell_1}|}{t_2} \cdots \binom{|V_{\ell_k} \setminus (V_{\ell_1} \cup \cdots \cup V_{\ell_{k-1}})|}{t_k},$$

$$\begin{aligned}\beta_{\vec{\sigma}}(\vec{\ell}) := \mathbb{P}[\ & I_1 \cup \cdots \cup I_k \subseteq \mathrm{Stars}(\boldsymbol{\varrho}) \text{ and} \\ & C_1(\boldsymbol{\varrho}) = \cdots = C_{\ell_1-1}(\boldsymbol{\varrho}) = 0 \text{ and } C_{\ell_1}(\boldsymbol{\varrho}\tau_1) = 1 \\ & C_{\ell_1+1}(\boldsymbol{\varrho}\sigma_1) = \cdots = C_{\ell_2-1}(\boldsymbol{\varrho}\sigma_1) = 0 \text{ and } C_{\ell_2}(\boldsymbol{\varrho}\sigma_1\tau_2) = 1 \text{ and} \\ & \quad \vdots \\ & C_{\ell_{k-1}+1}(\boldsymbol{\varrho}\sigma_1 \ldots \sigma_{k-1}) = \cdots = C_{\ell_k-1}(\boldsymbol{\varrho}\sigma_1 \ldots \sigma_{k-1}) = 0 \text{ and } C_{\ell_k}(\boldsymbol{\varrho}\sigma_1 \ldots \sigma_{k-1}\tau_k) = 1\ ].\end{aligned}$$

Note that the pair (ℓ_j, σ_j) determines both I_j (= $\mathrm{Dom}(\sigma_j)$) and τ_j (= the subclause of C_{ℓ_j} over I_j) for each $j \in [k]$. For this reason, we streamline notation by writing $\max_{\sigma_j}$ instead of $\max_{I_j,\sigma_j,\tau_j}$ and $\beta_{\vec{\sigma}}(\vec{\ell})$ instead of $\beta_{\vec{I},\vec{\sigma},\vec{\tau}}(\vec{\ell})$. Observe that $\alpha(\vec{\ell})$ is an upper bound on the number of choices of $\vec{\sigma}$ for a given $\vec{\ell}$.

Let $\boldsymbol{x} \in \{0,1\}^n$ be a uniform random completion of $\boldsymbol{\varrho}$ (i.e. a uniform random element of $\{0,1\}^n$ subject to $\boldsymbol{x}_i = \boldsymbol{\varrho}_i$ for all $i \in \mathrm{Dom}(\boldsymbol{\varrho})$). For a restriction $\pi \in \{0,1\}^J$, let $\boldsymbol{x}^\pi \in \{0,1\}^n$ denote $\boldsymbol{x}$ overwritten by π (i.e. $\boldsymbol{x}^\pi_i = \boldsymbol{x}_i$ for all $i \in [n] \setminus J$ and $\boldsymbol{x}^\pi_j = \pi_j$ for all $j \in J$). Using the independence of random variables $\mathrm{Stars}(\boldsymbol{\varrho})$ and $\boldsymbol{x}$, we have

$$\begin{aligned}\beta_{\vec{\sigma}}(\vec{\ell}) \leq \quad & \mathbb{P}[\ I_1 \cup \cdots \cup I_k \subseteq \mathrm{Stars}(\boldsymbol{\varrho}) \text{ and} \\ & \quad C_1(\boldsymbol{x}^{\tau_1\cdots\tau_k}) = \cdots = C_{\ell_1-1}(\boldsymbol{x}^{\tau_1\cdots\tau_k}) = 0 \text{ and } C_{\ell_1}(\boldsymbol{x}^{\tau_1\cdots\tau_k}) = 1 \text{ and} \\ & \quad C_{\ell_1+1}(\boldsymbol{x}^{\sigma_1\tau_2\cdots\tau_k}) = \cdots = C_{\ell_2-1}(\boldsymbol{x}^{\sigma_1\tau_2\cdots\tau_k}) = 0 \text{ and } C_{\ell_2}(\boldsymbol{x}^{\sigma_1\tau_2\cdots\tau_k}) = 1 \text{ and} \\ & \quad\ \vdots \\ & \quad C_{\ell_{k-1}+1}(\boldsymbol{x}^{\sigma_1\cdots\sigma_{k-1}\tau_k}) = \cdots = C_{\ell_k-1}(\boldsymbol{x}^{\sigma_1\cdots\sigma_{k-1}\tau_k}) = 0 \text{ and } C_{\ell_k}(\boldsymbol{x}^{\sigma_1\cdots\sigma_{k-1}\tau_k}) = 1\ ] \\ = (2p)^t\, & \mathbb{P}[\ \boldsymbol{x} \text{ extends } \tau_1 \ldots \tau_k \text{ (i.e., } \boldsymbol{x}^{\tau_1\cdots\tau_k} = \boldsymbol{x}) \text{ and} \\ & \quad C_1(\boldsymbol{x}^{\tau_1\cdots\tau_k}) = \cdots = C_{\ell_1-1}(\boldsymbol{x}^{\tau_1\cdots\tau_k}) = 0 \text{ and } C_{\ell_1}(\boldsymbol{x}^{\tau_1\cdots\tau_k}) = 1 \text{ and} \\ & \quad C_{\ell_1+1}(\boldsymbol{x}^{\sigma_1\tau_2\cdots\tau_k}) = \cdots = C_{\ell_2-1}(\boldsymbol{x}^{\sigma_1\tau_2\cdots\tau_k}) = 0 \text{ and } C_{\ell_2}(\boldsymbol{x}^{\sigma_1\tau_2\cdots\tau_k}) = 1 \text{ and} \\ & \quad\ \vdots \\ & \quad C_{\ell_{k-1}+1}(\boldsymbol{x}^{\sigma_1\cdots\sigma_{k-1}\tau_k}) = \cdots = C_{\ell_k-1}(\boldsymbol{x}^{\sigma_1\cdots\sigma_{k-1}\tau_k}) = 0 \text{ and } C_{\ell_k}(\boldsymbol{x}^{\sigma_1\cdots\sigma_{k-1}\tau_k}) = 1\ ] \\ = (2p)^t & \mu_{\vec{\sigma}}(\vec{\ell})\end{aligned}$$

where

$$\begin{aligned}\mu_{\vec{\sigma}}(\vec{\ell}) := \mathbb{P}[\ & C_1(\boldsymbol{x}) = \cdots = C_{\ell_1-1}(\boldsymbol{x}) = 0 \text{ and } C_{\ell_1}(\boldsymbol{x}) = 1 \text{ and} \\ & C_{\ell_1+1}(\boldsymbol{x}^{\sigma_1}) = \cdots = C_{\ell_2-1}(\boldsymbol{x}^{\sigma_1}) = 0 \text{ and } C_{\ell_2}(\boldsymbol{x}^{\sigma_1}) = 1 \text{ and} \\ & \quad \vdots \\ & C_{\ell_{k-1}+1}(\boldsymbol{x}^{\sigma_1\cdots\sigma_{k-1}}) = \cdots = C_{\ell_k-1}(\boldsymbol{x}^{\sigma_1\cdots\sigma_{k-1}}) = 0 \text{ and } C_{\ell_k}(\boldsymbol{x}^{\sigma_1\cdots\sigma_{k-1}}) = 1\ ].\end{aligned}$$

(Here we have used the fact that $C_{\ell_1}(\boldsymbol{x}) = C_{\ell_2}(\boldsymbol{x}^{\sigma_1}) = \cdots = C_{\ell_k}(\boldsymbol{x}^{\sigma_1\cdots\sigma_{k-1}}) = 1$ implies $\boldsymbol{x}^{\tau_1\cdots\tau_k} = \boldsymbol{x}$.) Combining (4) with the bound $\beta_{\vec{\sigma}}(\vec{\ell}) \leq (2p)^t \mu_{\vec{\sigma}}(\vec{\ell})$, we have

$$\mathbb{P}[\ \boldsymbol{\varrho} \text{ is } \vec{t}\text{-bad }] \leq (2p)^t \sum_{\ell_1} \max_{\sigma_1} \sum_{\ell_2} \max_{\sigma_2} \cdots \sum_{\ell_k} \max_{\sigma_k} \alpha(\vec{\ell})\mu_{\vec{\sigma}}(\vec{\ell}). \tag{5}$$

The next step in the proof rewrites (5) by replacing each $\sum_{\ell_j} \max_{\sigma_j}$ with $\max_{\sigma_j^*} \sum_{\ell_j}$ in the following manner. For $j \in [k]$, let L_j be the set of j-tuples $(\ell_1, \ldots, \ell_j)$ which extend to at least one k-tuple $\vec{\ell} = (\ell_1, \ldots, \ell_k)$ satisfying $1 \le \ell_1 < \cdots < \ell_k \le m$ and $|V_{\ell_i} \setminus (V_{\ell_1} \cup \cdots \cup V_{\ell_{i-1}})| \ge t_i$. Let I_j^* and σ_j^* range over functions on L_j mapping each $(\ell_1, \ldots, \ell_j) \in L_j$ to a choice of

$$I_j^*(\ell_1, \ldots, \ell_j) \in \binom{V_{\ell_j} \setminus (V_{\ell_1} \cup \cdots \cup V_{\ell_{j-1}})}{t_j} \quad \text{and} \quad \sigma_j^*(\ell_1, \ldots, \ell_j) \in \{0,1\}^{I_j^*(\ell_1, \ldots, \ell_j)}.$$

(Note: Since σ_j^* determines I_j^*, we simplify notation by indexing over $\vec{\sigma}^* = (\sigma_1^*, \ldots, \sigma_k^*)$ alone.) This allows us to rewrite (5) as

$$\mathbb{P}[\ \boldsymbol{\varrho} \text{ is } \vec{t}\text{-bad }] \le (2p)^t \max_{\vec{\sigma}^*} \sum_{\vec{\ell}} \alpha(\vec{\ell}) \mu_{\vec{\sigma}^*}(\vec{\ell}) \tag{6}$$

where

$$\begin{aligned}
\mu_{\vec{\sigma}^*}(\vec{\ell}) := \mathbb{P}[\ & C_1(\boldsymbol{x}) = \cdots = C_{\ell_1 - 1}(\boldsymbol{x}) = 0 \text{ and } C_{\ell_1}(\boldsymbol{x}) = 1 \text{ and} \\
& C_{\ell_1+1}(\boldsymbol{x}^{\sigma_1^*(\ell_1)}) = \cdots = C_{\ell_2 - 1}(\boldsymbol{x}^{\sigma_1^*(\ell_1)}) = 0 \text{ and } C_{\ell_2}(\boldsymbol{x}^{\sigma_1^*(\ell_1)}) = 1 \text{ and} \\
& C_{\ell_2+1}(\boldsymbol{x}^{\sigma_1^*(\ell_1)\sigma_2^*(\ell_1,\ell_2)}) = \cdots = C_{\ell_3 - 1}(\boldsymbol{x}^{\sigma_1^*(\ell_1)\sigma_2^*(\ell_1,\ell_2)}) = 0 \text{ and } C_{\ell_3}(\boldsymbol{x}^{\sigma_1^*(\ell_1)\sigma_2^*(\ell_1,\ell_2)}) = 1 \text{ and} \\
& \ \vdots \\
& C_{\ell_{k-1}+1}(\boldsymbol{x}^{\sigma_1^*(\ell_1)\ldots\sigma_{k-1}^*(\ell_1,\ldots,\ell_{k-1})}) = \cdots = C_{\ell_k - 1}(\boldsymbol{x}^{\sigma_1^*(\ell_1)\ldots\sigma_{k-1}^*(\ell_1,\ldots,\ell_{k-1})}) = 0 \\
& \qquad\qquad \text{and } C_{\ell_k}(\boldsymbol{x}^{\sigma_1^*(\ell_1)\ldots\sigma_{k-1}^*(\ell_1,\ldots,\ell_{k-1})}) = 1\ ].
\end{aligned}$$

For any fixed $\vec{\sigma}^*$, observe that the events defining $\mu_{\vec{\sigma}^*}(\vec{\ell})$ are mutually exclusive as $\vec{\ell}$ varies. Therefore, $\sum_{\vec{\ell}} \mu_{\vec{\sigma}^*}(\vec{\ell}) \le 1$. (Note: It is important here that σ_j^* is a function of $(\ell_1, \ldots, \ell_j) \in L_j$ alone and not the entire sequence $\vec{\ell} = (\ell_1, \ldots, \ell_k)$.)

We next turn to bounding $\alpha(\vec{\ell})$. First observe that

$$\begin{aligned}
\mu_{\vec{\sigma}^*}(\vec{\ell}) &\le \mathbb{P}[\ C_{\ell_1}(\boldsymbol{x}) = C_{\ell_2}(\boldsymbol{x}^{\sigma_1^*(\ell_1)}) = \cdots = C_{\ell_k}(\boldsymbol{x}^{\sigma_1^*(\ell_1)\ldots\sigma_{k-1}^*(\ell_1,\ldots,\ell_{k-1})}) = 1\ ] \\
&= \begin{cases} 2^{-|V_{\ell_1} \cup \cdots \cup V_{\ell_k}|} & \text{if } C_{\ell_1} \wedge C_{\ell_2}{\restriction}\sigma_1^*(\ell_1) \wedge \cdots \wedge C_{\ell_k}{\restriction}\sigma_1^*(\ell_1)\ldots\sigma_{k-1}^*(\ell_1,\ldots,\ell_{k-1}) \text{ is satisfiable,} \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

Therefore,

$$|V_{\ell_1} \cup \cdots \cup V_{\ell_k}| \le \log(1/\mu_{\vec{\sigma}^*}(\vec{\ell})).$$

It follows that

$$\begin{aligned}
\alpha(\vec{\ell}) &= 2^t \binom{|V_{\ell_1}|}{t_1} \binom{|V_{\ell_2} \setminus V_{\ell_1}|}{t_2} \cdots \binom{|V_{\ell_k} \setminus (V_{\ell_1} \cup \cdots \cup V_{\ell_{k-1}})|}{t_k} \\
&\le 2^t \binom{|V_{\ell_1}| + |V_{\ell_2} \setminus V_{\ell_1}| + |V_{\ell_k} \setminus (V_{\ell_1} \cup \cdots \cup V_{\ell_{k-1}})|}{t_1 + t_2 + \cdots + t_k} \\
&\le \left(\frac{2e|V_{\ell_1} \cup \cdots \cup V_{\ell_k}|}{t} \right)^t \\
&\le \left(\frac{2e \log(1/\mu_{\vec{\sigma}^*}(\vec{\ell}))}{t} \right)^t.
\end{aligned}$$

Combining this bound on $\alpha(\vec{\ell})$ with (6), we have

$$\mathbb{P}[\ \boldsymbol{\varrho} \text{ is } \vec{t}\text{-bad }] \leq \left(\frac{4ep}{\ln 2}\right)^t \max_{\vec{\sigma}^*} \sum_{\vec{\ell}} \mu_{\vec{\sigma}^*}(\vec{\ell}) \left(\frac{\ln(1/\mu_{\vec{\sigma}^*}(\vec{\ell}))}{t}\right)^t. \tag{7}$$

Since $\sum_{\vec{\ell}} \mu_{\vec{\sigma}^*}(\vec{\ell}) \leq 1$ and $\mu_{\vec{\sigma}^*}(\cdot)$ has support size $\leq \binom{m}{k}$ (i.e. the number of sequences $1 \leq \ell_1 < \cdots < \ell_k \leq m$), using Lemma 7 and the fact that $\binom{m}{k} \leq m^t$ (since $k \leq t$), we have

$$\sum_{\vec{\ell}} \mu_{\vec{\sigma}^*}(\vec{\ell}) \left(\frac{\ln(1/\mu_{\vec{\sigma}^*}(\vec{\ell}))}{t}\right)^t \leq \left(\frac{\ln(\binom{m}{k})}{t} + 2\right)^t \leq (\ln(e^2 m))^t.$$

Combining the above inequality with (7), we get the desired bound $\mathbb{P}[\ \boldsymbol{\varrho} \text{ is } \vec{t}\text{-bad }] \leq (4ep \log(e^2 m))^t$. □

Remark 12. We obtain a slightly better bound in Lemma 11 (and consequently in Theorem 2) by observing

- if $t \leq m/2$, then $\dfrac{\ln(\binom{m}{k})}{t} \leq \dfrac{\ln(\binom{m}{t})}{t} \leq \dfrac{\ln((em/t)^t)}{t} = \ln(em/t)$,
- if $t > m/2$, then $\dfrac{\ln(\binom{m}{k})}{t} \leq \dfrac{\ln(2^m)}{t} \leq \dfrac{m \ln(2)}{t} \leq \ln(4)$.

This leads to the bound

$$\mathbb{P}[\ \mathbf{R}_p \text{ is } \vec{t}\text{-bad }] \leq (4ep \log(e^2 \max\{\tfrac{em}{t}, 4\}))^t = O(p \log(\tfrac{m}{t} + 2))^t.$$

Note that this beats $O(p \log(m+1))^t$ for $t \geq m^{1-\Omega(1)}$. In particular, we get $O(p)^t$ for $t \geq m$.

Lemma 11 has the following corollary.

Corollary 13. $\mathbb{P}[\ \mathrm{CDT}(F\restriction \mathbf{R}_p) \text{ *has depth* } t\ ] \leq (8ep \log(e^2 m))^t$.

Proof. For any restriction ϱ such that $\mathrm{CDT}(F\restriction\varrho)$ has depth t, there exists $k \in \mathbb{N}_+$ and $\vec{t} \in \mathbb{N}_+^k$ with $t_1 + \cdots + t_k = t$ such that ϱ is $\vec{t}$-bad. The number of such pairs $(k, \vec{t})$ for a given t is exactly 2^{t-1}. Corollary 13 thus follows from Lemma 11 by a union bound. □

Theorem 2 (our switching lemma for m-clause DNFs) follows easily from Corollary 11 by an additional union bound.

Proof of Theorem 2. We will show

$$\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(F\restriction \mathbf{R}_p) \geq t\ ] \leq (16ep \log(e^2 m)) = O(p \log(m+1))^t.$$

We assume that $p \leq (16e \log(e^2 m))^{-1}$ and $t \geq 1$ (since the above inequality is trivial otherwise). By a union bound and Corollary 13, we have

$$\begin{aligned}
&\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(F\restriction \mathbf{R}_p) \geq t\ ] \\
&\qquad \leq \mathbb{P}[\ \mathrm{CDT}(F\restriction \mathbf{R}_p) \text{ has depth} \geq t\ ] \\
&\qquad \leq \sum_{i=0}^{\infty} \mathbb{P}[\ \mathrm{CDT}(F\restriction \mathbf{R}_p) \text{ has depth } t+i\ ] \\
&\qquad \leq \sum_{i=0}^{\infty} (8ep \log(e^2 m))^{t+i} \ \leq\ (8ep \log(e^2 m))^t \sum_{i=0}^{\infty} 2^{-i} \ \leq\ (16ep \log(e^2 m))^t.
\end{aligned}$$

□

4.1 Applications and Extensions of Theorem 2

We begin with the observation that Theorem 2 applies equally to m-term CNFs (by duality of DNFs and CNFs and invariance of DT under negations). As an aside, let us point out the proof of Theorem 2 implies the bound $\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(\widetilde{F}\restriction \mathbf{R}_p) \geq t\ ] \leq O(p\log(m+1))^t$ for the "first witness function" $\widetilde{F} : \{0,1\}^n \to [m+1]$ (similarly, proofs of Theorem 1 imply $\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(\widetilde{F}\restriction \mathbf{R}_p) \geq t\ ] \leq O(pw)^t$ for width-w DNFs F).

Since every depth-w decision tree is equivalent to both a 2^w-clause DNF and a 2^w-term CNF, Theorem 2 implies the following special case of Håstad's switching lemma (Theorem 1):

Corollary 14. *If f is a disjunction or conjunction of $2^{O(w)}$ many depth-w decision trees (and hence equivalent to a width-w DNF or CNF with $2^{O(w)}$ clauses), then $\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(f\restriction \mathbf{R}_p) \geq t\ ] = O(pw)^t$.*

In all applications of Håstad's switching lemma in circuit complexity that the author is aware of, Corollary 14 may be used instead. That is, the width-w DNFs and CNFs that arise in applications of Håstad's switching lemma are disjunctions and conjunctions of $2^{O(w)}$ many depth-w decision trees. For example, in the classic $2^{\Omega(n^{1/(d-1)})}$ lower bound on the depth-d AC^0 circuit size of PARITY, Håstad's switching lemma is applied (at each gate of the circuit) to disjunctions and conjunctions of at most s many decision trees of depth $O(\log s)$. Corollary 14 thus provides an alternative proof of an $2^{\Omega(n^{1/(d-1)})}$ lower bound for PARITY. (This equivalence of Theorems 1 and 2 for applications in circuit complexity justifies our use of the definite article in our title "an entropy proof of *the* switching lemma".)

One potential advantage of our switching lemma for m-clause DNFs is that its proof extends directly to a slightly broader class of random restrictions. We say that a random restriction $\boldsymbol{\rho}$ is *p-pseudorandom* if it satisfies

- $\mathbb{P}[\ I \subseteq \mathrm{Stars}(\boldsymbol{\rho})\ ] \leq p^{|I|}$ for all $I \subseteq [n]$,
- $\mathbb{P}[\ \boldsymbol{\rho}_i = 0 \mid i \in \mathrm{Dom}(\boldsymbol{\rho})\ ] = \mathbb{P}[\ \boldsymbol{\rho}_i = 1 \mid i \in \mathrm{Dom}(\boldsymbol{\rho})\ ] = \frac{1}{2}$ independently for all $i \in [n]$ (so that a uniform random completion $\boldsymbol{x}$ of $\boldsymbol{\rho}$ is uniformly distributed in $\{0,1\}^n$).

Corollary 15. *If F is an m-term DNF and $\boldsymbol{\rho}$ is a p-pseudorandom restriction, then*

$$\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(F\restriction \boldsymbol{\rho}) \geq t\ ] = O(p\log(m+1))^t.$$

The proof of Corollary 15 directly generalizes Theorem 2. (The key point is that the bound $\beta_{\vec{\sigma}}(\vec{\ell}) \leq (2p)^t \mu_{\vec{\sigma}}(\vec{\ell})$ in the proof of Lemma 11 applies to any p-pseudorandom $\boldsymbol{\rho}$.) In contrast, previous proofs of Håstad's switching lemma do not appear to readily extend to p-pseudorandom restrictions. This suggests that the entropy technique might be useful in obtaining new switching lemmas for other more general classes of random restrictions. The seemingly greater flexibility of the entropy technique might also be useful in the design of pseudorandom generators.

5 Criticality and Decision-Tree Size

For every boolean function f, the random variable $\mathsf{DT}_{\mathsf{depth}}(f\restriction \mathbf{R}_p)$ obeys an exponential tail bound for all sufficiently small $p > 0$. So far as I know, there is no name in the literature for the threshold value of p where an exponential tail bound takes hold. Let me offer:

Definition 16. A boolean function f is *p-critical* if $\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(f\restriction \mathbf{R}_p) \geq t\ ] \leq \exp(-t)$ for all $t \in \mathbb{N}$.

Note that if f depends on n variables, then it is $1/en$-critical, as $\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(f\restriction \mathbf{R}_{1/en}) \geq t\ ] \leq \mathbb{P}[\ \mathbf{Bin}(n, 1/en) \geq t\ ] \leq \exp(-t)$. Thus, every boolean function of finitely many variables is p-critical for some $p > 0$. The next two propositions give key properties of p-critical functions. Psroposition 17 in particular, though simple and conceivably folklore, makes an useful connection between criticality and decision-tree size (and, by extension, satisfiable algorithms).

Proposition 17. *Every p-critical boolean function of n variables has decision-tree size $\leq 20 \cdot 2^{(1-p)n}$.*

Proof. Suppose $f : \{0,1\}^n \to \{0,1\}$ is p-critical. Let $\boldsymbol{S}$ be a $1-p$-binomial random subset of $[n]$ (with density function $\mathbb{P}[\ \boldsymbol{S} = S\ ] = (1-p)^{|S|}p^{n-|S|}$). Let $\boldsymbol{\varrho} \in \{0,1\}^{\boldsymbol{S}}$ be a uniform random restriction with domain $\boldsymbol{S}$. Note that $\boldsymbol{\varrho}$, on its own, is a p-random restriction.

Summarizing the proof: we obtain a decision tree for f by sampling $\boldsymbol{S}$, querying all variables in $\boldsymbol{S}$, and then appending the optimal decision tree for $f\restriction\varrho$ for each $\varrho \in \{0,1\}^{\boldsymbol{S}}$. We will show that the resulting decision tree has size $\leq 20 \cdot 2^{(1-p)n}$ with nonzero probability. By the magic of the probabilistic method, we conclude that $\mathsf{DT}_{\mathsf{size}}(f) \leq 20 \cdot 2^{(1-p)n}$.

First, we observe that, for any fixed S,

$$\mathsf{DT}_{\mathsf{size}}(f) \leq \sum_{\varrho \in \{0,1\}^S} 2^{\mathsf{DT}_{\mathsf{depth}}(f\restriction\varrho)} = 2^{|S|} \mathop{\mathbb{E}}_{\boldsymbol{\varrho}\in\{0,1\}^S}[\ 2^{\mathsf{DT}_{\mathsf{depth}}(f\restriction\boldsymbol{\varrho})}\ ]. \tag{8}$$

Since every median of $\mathbf{Bin}(n,p)$ is at least $\lfloor pn \rfloor$, we have

$$\mathbb{P}[\ |\boldsymbol{S}| > \lceil (1-p)n \rceil\ ] = \mathbb{P}[\ \mathbf{Bin}(n, 1-p) > n - \lfloor pn \rfloor\ ] = \mathbb{P}[\ \mathbf{Bin}(n,p) < \lfloor pn \rfloor\ ] \leq \frac{1}{2}. \tag{9}$$

We now have

$$\begin{aligned}
\mathop{\mathbb{P}}_{\boldsymbol{S}}[\ \mathsf{DT}_{\mathsf{size}}(f) > 20 \cdot 2^{(1-p)n}\ ] &\leq \mathop{\mathbb{P}}_{\boldsymbol{S}}\left[\ 2^{|\boldsymbol{S}|} \mathop{\mathbb{E}}_{\boldsymbol{\varrho}\in\{0,1\}^{\boldsymbol{S}}}[\ 2^{\mathsf{DT}_{\mathsf{depth}}(f\restriction\boldsymbol{\varrho})}\ ] > 20 \cdot 2^{(1-p)n}\ \right] \quad \text{(by (8))} \\
&\leq \mathop{\mathbb{P}}_{\boldsymbol{S}}\left[\ \left(2^{|\boldsymbol{S}|} > 2^{(1-p)n+1}\right) \vee \left(\mathop{\mathbb{E}}_{\boldsymbol{\varrho}\in\{0,1\}^{\boldsymbol{S}}}[\ 2^{\mathsf{DT}_{\mathsf{depth}}(f\restriction\boldsymbol{\varrho})}\ ] > 10\right)\ \right] \\
&\leq \mathop{\mathbb{P}}_{\boldsymbol{S}}[\ |\boldsymbol{S}| > \lceil (1-p)n \rceil\ ] + \mathop{\mathbb{P}}_{\boldsymbol{S}}\left[\ \mathop{\mathbb{E}}_{\boldsymbol{\varrho}\in\{0,1\}^{\boldsymbol{S}}}[\ 2^{\mathsf{DT}_{\mathsf{depth}}(f\restriction\boldsymbol{\varrho})}\ ] > 10\ \right] \\
&\leq \frac{1}{2} + \frac{1}{10}\,\mathbb{E}[\ 2^{\mathsf{DT}_{\mathsf{depth}}(f\restriction\mathbf{R}_p)}\ ] \quad \text{(by (9) and Markov's inequality)} \\
&= \frac{1}{2} + \frac{1}{10}\sum_{t=0}^{\infty} 2^t \cdot \underbrace{\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(f\restriction\mathbf{R}_p) = t\ ]}_{\leq \exp(-t) \text{ by } p\text{-criticality of } f} \\
&= \frac{1}{2} + \frac{1}{10} \cdot \frac{1}{1-(2/e)} \\
&< 1.
\end{aligned}$$

It follows that $\mathsf{DT}_{\mathrm{size}}(f) \leq 20 \cdot 2^{(1-p)n}$. □

Proposition 18. *If f is p-critical, then $\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(f\restriction\mathbf{R}_q) \geq t\ ] = O(q/p)^t$ for all $q \in [0,1]$ and $t \in \mathbb{N}$.*

Proof. We assume that $q \in [0,p]$ and $t \geq 1$ (since the bound is trivial otherwise). Generate $\mathbf{R}_q$ as the composition of a random restriction $\boldsymbol{\varrho}_1 \sim \mathbf{R}_p$ (over the variables of f) and $\boldsymbol{\varrho}_2 \sim \mathbf{R}_{q/p}$ (over the variables of $f\restriction\boldsymbol{\varrho}_1$). We have

$$\begin{aligned}
&\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(f\restriction\mathbf{R}_q) \geq t\ ] \\
&\quad= \mathop{\mathbb{E}}_{\boldsymbol{\varrho}_1}\Big[\ \mathop{\mathbb{P}}_{\boldsymbol{\varrho}_2}[\ \mathsf{DT}_{\mathsf{depth}}((f\restriction\boldsymbol{\varrho}_1)\restriction\boldsymbol{\varrho}_2) \geq t\ ]\ \Big] \\
&\quad= \sum_{i=0}^{\infty} \underbrace{\mathop{\mathbb{P}}_{\boldsymbol{\varrho}_1}[\ \mathsf{DT}_{\mathsf{depth}}(f\restriction\boldsymbol{\varrho}_1) = t+i\ ]}_{\leq \exp(-t-i) \text{ by } p\text{-criticality of } f} \cdot \mathop{\mathbb{E}}_{\boldsymbol{\varrho}_1}\Big[\ \underbrace{\mathop{\mathbb{P}}_{\boldsymbol{\varrho}_2}[\ \mathsf{DT}_{\mathsf{depth}}((f\restriction\boldsymbol{\varrho}_1)\restriction\boldsymbol{\varrho}_2) \geq t\ ]}_{\leq (2eq(t+i)/pt)^t \text{ by Cor. 21}}\ \Big|\ \mathsf{DT}_{\mathsf{depth}}(f\restriction\boldsymbol{\varrho}_1) = t+i\ \Big] \\
&\quad\leq (4eq/p)^t \cdot \sum_{i=0}^{\infty} \exp(-t-i) \cdot \big(\underbrace{(t+i)/2t}_{\leq \exp(i/2t)}\big)^t \\
&\quad\leq (4q/p)^t \cdot \sum_{i=0}^{\infty} \exp(-i/2) \\
&\quad= O(q/p)^t.
\end{aligned}$$

□

In light of Proposition 18, Theorems 1 and 2 are equivalent to the statements that every width-w DNF is $1/O(w)$-critical and every m-clause DNF is $1/O(\log(m+1))$-critical. The other main result of this paper, Theorem 3, is a combination of Propositions 17 and 18 with the following

Theorem 19 (Criticality of AC^0 Circuits)**.** *For all $d \geq 2$, every boolean function computed by an AC^0 circuit of depth d and size s is p-critical for $p = 1/O(\log s)^{d-1}$.*

Note that Theorem 2 (i.e., $1/O(\log(m+1))$-criticality of m-clause DNFs) is precisely the case $d = 2$ of Theorem 19. However, our proof of Theorem 2 (in the next section) does not involve the entropy argument of §4. Rather, we use a combination of Håstad's switching lemma and Håstad's recent "multi-switching lemma" [6], which was originally devised to obtain tight correlation bounds between AC^0 circuits and PARITY.[3] It would be interesting if one could prove Theorem 19 by an extension of the entropy argument in §4, or via a bound on the criticality of conjunctions of p-critical functions (see the "criticality question" in §7).

6 Proof of Theorem 19

We begin with a review (and mild reformulation) of Håstad's switching and multi-switching lemmas, as well as the even more basic shrinkage lemma for decision trees.

6.1 Decision-Tree Shrinkage

For a decision tree T and a restriction ϱ, let $T\restriction\varrho$ be the syntactically restricted decision tree (defined in the obvious way). We will require both of the following "syntactic" and "semantic" versions of the decision-tree shrinkage lemma.

[3] Roughly speaking, for width-w DNFs with $2^{O(w)}$ clauses, the switching lemma is effective for $t \leq w$, while the multi-switching lemma is effective for $t \geq w$. Our use of the switching and multi-switching lemmas in Appendix 6 is very similar to their use by Tal [12] in bounding the Fourier spectrum of AC^0 circuits.

Lemma 20 (Syntactic Decision-Tree Shrinkage Lemma)**.** *If T is a depth-k decision tree, then*

$$\mathbb{P}[\ T\restriction \mathbf{R}_p \text{ has depth} \geq \ell\ ] \leq (2epk/\ell)^\ell.$$

Proof. For any decision tree T, let random variable $\boldsymbol{Q}(T) \in \mathbb{N}$ be the number of variables queried by T on a uniform random input. This random variable has density function

$$\mathbb{P}[\ \boldsymbol{Q}(T) = \ell\ ] = 2^{-\ell} \cdot \#\{\text{leaves of } T \text{ at distance } \ell \text{ from the root}\}.$$

Suppose T has depth k. Without loss of generality, assume that no variable is queried more than once on any branch of T. Observe that random variables $\boldsymbol{Q}(T\restriction \mathbf{R}_p)$ and $\mathbf{Bin}(\boldsymbol{Q}(T), p)$ are identically distributed.

The lemma is proved by the following calculation:

$$\begin{aligned}
\mathbb{P}[\ T\restriction \mathbf{R}_p \text{ has depth} \geq \ell\ ] &= \mathop{\mathbb{P}}_{\mathbf{R}_p}[\ \mathop{\mathbb{P}}_{\boldsymbol{Q}(T\restriction \mathbf{R}_p)}[\ \boldsymbol{Q}(T\restriction \mathbf{R}_p) \geq \ell\ ] \geq 2^{-\ell}\ ] \\
&\leq 2^\ell\, \mathbb{P}[\ \boldsymbol{Q}(T\restriction \mathbf{R}_p) \geq \ell\ ] \quad \text{(Markov's inequality)} \\
&= 2^\ell\, \mathbb{P}[\ \mathbf{Bin}(\boldsymbol{Q}(T), p) \geq \ell\ ] \\
&\leq 2^\ell\, \mathbb{P}[\ \mathbf{Bin}(k, p) \geq \ell\ ] \ \leq\ (2p)^\ell \tbinom{k}{\ell} \ \leq\ (2epk/\ell)^\ell. \qquad \square
\end{aligned}$$

Corollary 21 (Semantic Decision-Tree Shrinkage Lemma)**.** *If f is a boolean function with decision-tree depth k, then*

$$\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(f\restriction \mathbf{R}_p) \geq \ell\ ] \leq (2epk/\ell)^\ell.$$

6.2 Håstad's Switching and Multi-Switching Lemmas

For parameters $d, k, s, t \in \mathbb{N}$, we speak of following classes (with respect to a common fixed set of variables, w.l.o.g. $[n]$):

- $\mathsf{DT}(k)$ is the class of depth-k decision trees.
- $\mathsf{CKT}(d, s)$ is the class of single-output AC^0 circuits of depth d and size s. If $s = s_1 + \cdots + s_d$ where $s_1, \ldots, s_{d-1} \geq 1$ and $s_d = 1$, let $\mathsf{CKT}(d; s_1, \ldots, s_d)$ denote the subclass of circuits in $\mathsf{CKT}(d, s)$ which have s_i depth-i subcircuits for all $i \in \{1, \ldots, d\}$.
- $\mathsf{CKT}(d, s) \circ \mathsf{DT}(k)$ is the class of circuits in $\mathsf{CKT}(d, s)$ whose inputs are labeled by decision trees in $\mathsf{DT}(t)$.
- $\mathsf{DT}(t) \circ \mathsf{CKT}(d, s) \circ \mathsf{DT}(k)$ is the class of depth-t decision trees, whose leaves are labeled by elements of $\mathsf{CKT}(d, s) \circ \mathsf{DT}(k)$.

(Recall that *circuit size* is the number of gates; depth-1 circuits have size 1; depth-0 circuits have size 0.) Note the following edge cases:

$$\begin{aligned}
\mathsf{CKT}(0, 0) = \mathsf{DT}(1) &= \{\text{literals and constants}\}, \\
\mathsf{CKT}(d, s) = \mathsf{CKT}(d, s) \circ \mathsf{DT}(1) &= \mathsf{DT}(0) \circ \mathsf{CKT}(d, s) \circ \mathsf{DT}(1).
\end{aligned}$$

We say that a boolean function f belongs to one of these classes if f is computed by an object in the class.

We next state Håstad's switching lemma [5] and multi-switching lemma [6] in the form that they are used in application to AC^0 circuits.

Lemma 22 (Håstad's Switching Lemma [5] + Union Bound)**.** *If $d \geq 1$ and $f \in \mathsf{CKT}(d; s_1, \ldots, s_d) \circ \mathsf{DT}(k)$, then*

$$\mathbb{P}[\ f\restriction \mathbf{R}_p \notin \mathsf{CKT}(d-1; s_2, \ldots, s_d) \circ \mathsf{DT}(t-1)\ ] \leq s_1(5pk)^t.$$

Proof. Consider the $\mathsf{CKT}(d; s_1, \ldots, s_d) \circ \mathsf{DT}(k)$ circuit which computes f. Each bottom-level gate is equivalent to a width-k DNF or CNF formula. The switching lemma (Theorem 1) implies that under the random restriction $\mathbf{R}_p$, each of these DNFs and CNFs lies in the class $\mathsf{DT}(t-1)$ with probability $\geq 1 - (5pk)^t$. The lemma follows by taking a union bound over the s_1 bottom-level gates. □

Lemma 23 (Håstad's Multi-Switching Lemma [6])**.** *If $d \geq 1$ and $f \in \mathsf{CKT}(d; s_1, \ldots, s_d) \circ \mathsf{DT}(k)$ and $\ell \geq \log s_1 + 1$, then*

$$\mathbb{P}[\ f\restriction \mathbf{R}_p \notin \mathsf{DT}(t-1) \circ \mathsf{CKT}(d-1; s_2, \ldots, s_d) \circ \mathsf{DT}(\ell)\ ] \leq s_1(50pk)^t.$$

This natural reformulation of the multi-switching lemma is due to Prahladh Harsha and Srikanth Srinivasan (personal communication). Håstad's originally devised this result in [6] in order to obtain nearly optimal correlation bounds between AC^0 circuits and PARITY. Impagliazzo, Matthews and Paturi [7] independently obtained a similar multi-switching lemma, which also gives nearly optimal correlation bounds between AC^0 circuits and PARITY (and which are in fact even better for almost-linear size $s \leq n^{1+o(1)}$).

6.3 Combined Multi-Switching Lemma

The main ingredient for our proof of Theorem 19 is the following lemma, which combines Håstad's multi-switching lemma with the syntactic decision-tree shrinkage lemma.

Lemma 24 (Combined Multi-Switching Lemma)**.** *If $d, t \geq 1$ and $f \in \mathsf{DT}(t-1) \circ \mathsf{CKT}(d; s_1, \ldots, s_d) \circ \mathsf{DT}(k)$ and $\ell \geq \log s_1 + 1$, then*

$$\mathbb{P}[\ f\restriction \mathbf{R}_p \notin \mathsf{DT}(t-1) \circ \mathsf{CKT}(d-1; s_2, \ldots, s_d) \circ \mathsf{DT}(\ell)\ ] \leq s_1(200pk)^{t/2}.$$

Observe that Lemma 24 involves a weaker hypothesis than Lemma 23 (f is assumed to lie in a large class). It bounds the probability of the same event, but gives a weaker bound ($s_1(200pk)^{t/2}$ instead of $s_1(50pk)^t$). The advantage of Lemma 24 is that it is suited to induction on d.

Proof. Suppose f is computed by a depth $t-1$ decision tree T, each of whose leaves λ is labeled by a circuit $C_\lambda \in \mathsf{CKT}(d, s, m) \circ \mathsf{DT}(k)$. Consider events

$$\mathcal{A} \stackrel{\text{def}}{\iff} T\restriction \mathbf{R}_p \text{ has depth } \leq \lceil t/2 \rceil - 1,$$
$$\mathcal{B} \stackrel{\text{def}}{\iff} C_\lambda\restriction \mathbf{R}_p \in \mathsf{DT}(\lceil t/2 \rceil - 1) \circ \mathsf{CKT}(d-1; s_2, \ldots, s_d) \circ \mathsf{DT}(\ell) \text{ for every leaf } \lambda \text{ of } T.$$

Note the implication

$$\mathcal{A} \wedge \mathcal{B} \implies f\restriction \mathbf{R}_p \in \mathsf{DT}(t-1) \circ \mathsf{CKT}(d-1; s_2, \ldots, s_d) \circ \mathsf{DT}(\log s + 1).$$

By Lemma 20 (the syntactic decision-tree shrinkage lemma), we have

$$\begin{aligned}\mathbb{P}[\ \neg\mathcal{A}\ ] &= \mathbb{P}[\ T\restriction\mathbf{R}_p \text{ has depth} \geq \lceil t/2\rceil\ ] \\ &\leq (2ep(t-1)/\lceil t/2\rceil)^{\lceil t/2\rceil} \\ &\leq (4ep)^{t/2}.\end{aligned}$$

By Lemma 23 (the multi-switching lemma) and a union bound, we have

$$\begin{aligned}\mathbb{P}[\ \neg\mathcal{B}\ ] &\leq \sum_{\lambda} \mathbb{P}[\ C_\lambda\restriction\mathbf{R}_p \notin \mathsf{DT}(\lceil t/2\rceil - 1) \circ \mathsf{CKT}(d-1; s_2, \ldots, s_d) \circ \mathsf{DT}(\ell)\ ] \\ &\leq \sum_{\lambda} s_1(50pk)^{\lceil t/2\rceil} \\ &\leq 2^{t-1} s_1(50pk)^{\lceil t/2\rceil}.\end{aligned}$$

Putting things together, we have

$$\begin{aligned}\mathbb{P}[\ f\restriction\mathbf{R}_p \notin \mathsf{DT}(t-1) \circ \mathsf{CKT}(d-1; s_2, \ldots, s_d) \circ \mathsf{DT}(\log s + 1)\ ] &\leq \mathbb{P}[\ \neg\mathcal{A}\ ] + \mathbb{P}[\ \neg\mathcal{B}\ ] \\ &\leq (4ep)^{t/2} + 2^{t-1} s_1(50pk)^{t/2} \\ &\leq \tfrac{1}{2}(16ep)^{t/2} + \tfrac{1}{2} s_1(200pk)^{t/2} \\ &\leq s_1(200pk)^{t/2}. \qquad \square\end{aligned}$$

We are finally ready to prove Theorem 19. The proof involves a similar use of the switching and multi-switching lemmas as in Håstad [5, 6] and Tal [12]. The only difference is our use of Lemma 24 (the combined multi-switching lemma) to deal with the outer decision tree at each stage of the restriction.

Proof of Theorem 19. Let C be a circuit of depth d and size s (where $d, s \geq 2$ without loss of generality), which computes a boolean function f. We wish to show that f is p-critical for $p = 1/O(\log s)^{d-1}$, that is,

$$\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(f\restriction\mathbf{R}_p) \geq t\ ] \leq \exp(-t)$$

for all $t \in \mathbb{N}$.

The case $t = 0$ is trivial. For the case $1 \leq t \leq \log s$, we will use Lemma 22 (the switching lemma + union bound) in the completely standard way. For the case $t \geq \log s$, we will use Lemma 24 (our "combined multi-switching lemma").

First, we fix some parameters. For $i \in \{1, \ldots, d\}$, let s_i be the number of depth-i subcircuits of C. Note that $s = s_1 + \cdots + s_d$ and $s_d = 1$. Let

$$\ell := \lceil \log s\rceil + 1, \quad p := \frac{1}{12800^{d+1}\ell^{d-1}}, \quad \text{and} \quad p_i = \frac{1}{12800^{i}\ell^{d-1}} \text{ for } i \in \{1, \ldots, d\}.$$

Note that $p = O(\log s)^d$ (as required) and $p_1 = p/p_d = 1/12800$ and $p_i/p_{i-1} = 1/12800\ell$ for all $i \in \{2, \ldots, d\}$.

Small t case: $1 \le t \le \log s$**.**

For $i \in \{1, \dots, d-1\}$, let $\mathcal{A}_i$ denote the event that $\mathsf{DT}_{\mathsf{depth}}(g\restriction \mathbf{R}_{p_i}) \le \ell$ for all functions g computed by depth-i subcircuits of C. By Lemma 22, we have

$$\mathbb{P}[\ \neg\mathcal{A}_1\ ] \le s_1(5p_1)^\ell = s_1(1/2560)^\ell.$$

Again by Lemma 22, we have

$$\mathbb{P}[\ \neg\mathcal{A}_2 \mid \mathcal{A}_1\ ] \le s_2(5(p_2/p_1)\ell)^\ell = s_2(1/2560)^\ell.$$

Here we view $\mathbf{R}_{p_2}$ as the composition of $\mathbf{R}_{p_1}$ (over the variables of f) and $\mathbf{R}_{p_2/p_1}$ (over the free variables of $\mathbf{R}_{p_1}$).

Similarly, for all $i \in \{2, \dots, d-1\}$, we have

$$\mathbb{P}[\ \neg\mathcal{A}_i \mid \mathcal{A}_1 \wedge \cdots \wedge \mathcal{A}_{i-1}\ ] \le s_i(1/2560)^\ell.$$

Therefore,

$$\begin{aligned}
\mathbb{P}[\ \neg\mathcal{A}_{d-1}\ ] &\le \textstyle\sum_{i=1}^{d-1} \mathbb{P}[\ \neg\mathcal{A}_i \mid \mathcal{A}_1 \wedge \cdots \wedge \mathcal{A}_{i-1}\ ] \\
&\le (s_1 + \cdots + s_{d-1})(1/2560)^\ell \\
&= (s-1)(1/2560)^\ell \\
&\le (1/1280)^\ell \quad (\text{since } \ell > \log s) \\
&\le (1/1280)^t \quad (\text{since } \ell > t).
\end{aligned}$$

By a final application of Lemma 22, we have

$$\begin{aligned}
\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(f\restriction \mathbf{R}_p) \ge t \mid \mathcal{A}_{d-1}\ ] &\le (5(p/p_{d-1})\ell)^t \\
&= (1/32768000)^t.
\end{aligned}$$

Combining the above inequalities, we get the desired bound

$$\begin{aligned}
\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(f\restriction \mathbf{R}_p) \ge t\ ] &\le \mathbb{P}[\ \neg\mathcal{A}_{d-1}\ ] + \mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(f\restriction \mathbf{R}_p) \ge t \mid \mathcal{A}_{d-1}\ ] \\
&\le (1/1280)^t + (1/32768000)^t \\
&< \exp(-t).
\end{aligned}$$

(The final inequality is easily shown to hold for all $t \ge 1$.)

Large t case: $t \ge \log s$**.**

Initially, we have $f \in \mathsf{CKT}(d; s_1, \dots, s_d) \circ \mathsf{DT}(1)$.

For $i \in \{1, \dots, d\}$, let $\mathcal{B}_i$ be the event

$$\mathcal{B}_i \iff^{\text{def}} f\restriction \mathbf{R}_{p_i} \in \mathsf{DT}(t-1) \circ \mathsf{CKT}(d-i; s_{i+1}, \dots, s_d) \circ \mathsf{DT}(\ell).$$

In particular, note that

$$\mathcal{B}_d \iff f\restriction \mathbf{R}_{p_d} \in \mathsf{DT}(t+\ell-1)$$

since $\mathsf{DT}(t-1) \circ \mathsf{CKT}(0,0) \circ \mathsf{DT}(\ell) = \mathsf{DT}(t+\ell-1)$.

By Lemma 23 (the multi-switching lemma), we have

$$\mathbb{P}[\ \neg\mathcal{B}_1\ ] \leq s_1(50p_1)^t = s_1(1/256)^t.$$

Next, for all $i = 2, \ldots, d$, by Lemma 24 (the combined multi-switching lemma) we have

$$\mathbb{P}[\ \neg\mathcal{B}_i \mid \mathcal{B}_1 \wedge \cdots \wedge \mathcal{B}_{i-1}\ ] \leq s_i(200(p_i/p_{i-1})\ell)^{t/2} = s_i(1/64)^{t/2} = s_i(1/8)^t.$$

Therefore,

$$\begin{aligned}
\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(f\restriction \mathbf{R}_{p_d}) \geq t+\ell\ ] &= \mathbb{P}[\ \neg\mathcal{B}_d\ ] \\
&\leq \textstyle\sum_{i=1}^{d} \mathbb{P}[\ \neg\mathcal{B}_i \mid \mathcal{B}_1 \wedge \cdots \wedge \mathcal{B}_{i-1}\ ] \\
&\leq s_1(1/256)^t + (s_2 + \cdots + s_d)(1/8)^t \\
&\leq s(1/8)^t \\
&\leq s(1/8)^{\frac{1}{3}\log s + \frac{2}{3}t} \qquad (\text{since } t \geq \log s) \\
&= (1/4)^t.
\end{aligned}$$

As a last step, we apply Lemma 21 (the semantic decision-tree shrinkage lemma) to get

$$\begin{aligned}
\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(f\restriction \mathbf{R}_p) \geq t \mid \mathsf{DT}_{\mathsf{depth}}(f\restriction \mathbf{R}_{p_d}) \leq t+\ell-1\ ] &\leq (2e(p/p_{d+1})t/(t+\ell-1))^t \\
&\leq (e/3200)^t
\end{aligned}$$

using $p/p_{d+1} = 1/12800$ and $t + \ell - 1 = t + \lceil \log s \rceil \geq 2t$.

Putting these inequalities together, we get the desired bound

$$\begin{aligned}
\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(f\restriction \mathbf{R}_p) \geq t\ ] &\leq \mathbb{P}[\ f\restriction \mathbf{R}_{p_d} \geq t+\ell\ ] + \mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(f\restriction \mathbf{R}_p) \geq t \mid f\restriction \mathbf{R}_{p_d} \leq t+\ell-1\ ] \\
&\leq (1/4)^t + (e/3200)^t \\
&\leq \exp(-t).
\end{aligned}$$

□

7 Open Questions

Prove that AC^0 formulas F of depth d and size s are $1/O(\frac{1}{d}\log s)^{d-1}$-critical. A result of the author in [10] implies that F satisfies

$$\mathbb{P}[\ \mathsf{DT}_{\mathsf{depth}}(F\restriction \mathbf{R}_p) \geq t\ ] \leq \exp(-t) \text{ where } p = 1/O(\tfrac{1}{d}\log s)^{d-1} \tag{10}$$

for all $t \leq O(\log s)$. To show that F is p-critical, it suffices to extend (10) to $t \geq \Omega(\log s)$. This would be interesting, at it implies a better bound on decision-tree size and, as a corollary (assuming a randomized procedure for obtaining the decision tree), a faster randomized SAT algorithm for AC^0 formulas vis-à-vis AC^0 circuits.

A more conventional entropy argument. Our proof of Theorem 2 relies on Lemma 7 involving the entropy-like quantity $\sum_i \mu_i \log(1/\mu_i)^t$. Is there an alternative (information-theoretic) proof of Theorem 2 that uses the more conventional Shannon entropy?

Criticality question. Suppose boolean functions $f_1, \ldots, f_m$ are *hereditarily p-critical*, meaning that every subfunction $f_i \restriction \varrho$ is p-critical (for all $i \in [m]$ and restriction ϱ). Is the function $f_1 \wedge \cdots \wedge f_m$ necessarily $p/O(\log(m+1))$-critical? If so, note that this directly implies Theorem 19.

Acknowledgements

I am grateful to Or Meir for helpful comments on the switching lemma proof. Theorem 19 on the criticality of AC^0 circuits emerged from conversations with Prahladh Harsha, Rahul Santhanam, Srikanth Srinivasan and Avishay Tal. I thank Ian Mertz and Toni Pitassi as well for helpful discussions. Finally, I thank Shrikanth Srinivasan, Siddharth Bhandari and Tulasi Molli for suggesting an improvement to the proof of Lemma 11 (replacing an application of the AM-GM inequality with a simpler combinatorial inequality).

References

[1] Kazuyuki Amano. Tight bounds on the average sensitivity of k-CNF. *Theory of Computing*, 7(1):45–48, 2011.

[2] Paul Beame. A switching lemma primer. Technical report, Technical Report UW-CSE-95-07-01, Department of Computer Science and Engineering, University of Washington, 1994.

[3] Paul Beame, Russell Impagliazzo, and Srikanth Srinivasan. Approximating AC^0 by small height decision trees and a deterministic algorithm for $\#\mathrm{AC}^0$-SAT. In *Computational Complexity (CCC), 2012 IEEE 27th Annual Conference on*, pages 117–125. IEEE, 2012.

[4] Ravi B Boppana. The average sensitivity of bounded-depth circuits. *Information processing letters*, 63(5):257–261, 1997.

[5] Johan Håstad. Almost optimal lower bounds for small depth circuits. In *Proceedings of the 18th Annual ACM Symposium on Theory of Computing*, pages 6–20. ACM, 1986.

[6] Johan Håstad. On the correlation of parity and small-depth circuits. *SIAM Journal on Computing*, 43(5):1699–1708, 2014.

[7] Russell Impagliazzo, William Matthews, and Ramamohan Paturi. A satisfiability algorithm for AC^0. In *Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 961–972. SIAM, 2012.

[8] Nathan Keller and Noam Lifshitz. Approximation of biased boolean functions of small total influence by DNF's. *arXiv preprint arXiv:1703.10116*, 2017.

[9] Alexander A Razborov. An equivalence between second order bounded domain bounded arithmetic and first order bounded arithmetic. 1993.

[10] Benjamin Rossman. The average sensitivity of bounded-depth formulas. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 424–430. IEEE, 2015.

[11] Dominik Scheder and Li-Yang Tan. On the average sensitivity and density of k-CNF formulas. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 683–698. Springer, 2013.

[12] Avishay Tal. Tight bounds on the Fourier spectrum of AC^0. In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 21, page 174, 2014.

[13] Patrick Traxler. Variable influences in conjunctive normal forms. In *Theory and Applications of Satisfiability Testing-SAT 2009: 12th International Conference, SAT 2009, Swansea, UK, June 30-July 3, 2009. Proceedings*, volume 5584, page 101. Springer, 2009.