Electronic Colloquium on Computational Complexity, Report No. 41 (2016)

An average-case depth hierarchy theorem for higher depths

Johan Håstad*
KTH - Royal Institute of Technology

March 17, 2016

Abstract

We extend the recent hierarchy results of Rossman, Servedio and Tan [2] to any $d \le \frac{c \log n}{\log \log n}$ for an explicit constant c.

To be more precise, we prove that for any such d there is a function F_d that is computable by a read-once formula of depth d but such that any circuit of depth $d-1$ and size at most $2^{O(n^{1/5d})}$ agrees with F_d on a fraction at most $\frac{1}{2} + O(n^{-1/8d})$ of inputs.

1 Introduction

This technical note is an extension of the work of Rossman, Servedio and Tan [2]. Their result is that there is a constant c such that for $d \le \frac{c\sqrt{\log n}}{\log \log n}$ there is a function F_d that is computed by a read-once formula of depth d but such that any circuit of size at most $\exp(n^{\frac{1}{6(d-1)}})$ and depth $d-1$ agrees with F_d on a fraction of inputs that is at most $\frac{1}{2} + n^{-\Omega(1/d)}$. We extend their result by allowing d to be as large as $\frac{\log n}{\log \log n}$ and in particular we prove the following theorem.

Theorem 1.1 *For sufficiently large n and $d \le \frac{\log n}{\log \log n}$, then there is a Boolean function F_d, depending on n Boolean inputs which is computable by a read-once circuit of depth d such that for any circuit C of depth $d-1$ and size at most $2^{n^{1/5(d-1)}}$ we have*

$$Pr[F_d(x) = C(x)] \le \frac{1}{2} + O(n^{-\Omega(1/8d)}).$$

The theorem is only really interesting for $d \le \frac{\log n}{5 \log \log n}$ as the bound on the size of C is sublinear for larger values of d. For the record let us state that we

*Research supported the Swedish Research Council, part of this work was done while visiting the Simons Institute in Berkeley.

ISSN 1433-8092

have not tried to optimize the value of the constants 5 and 8, but even being more careful, we do not believe they can be pushed below 2.

In fact, as was also done in [2], we prove a lower bound for a slightly larger class of circuits including those of depth d and small bottom fanin and even depth $d+1$ circuits with small bottom fanin if the output gate is restricted to be the other type compared to the output gate of F_d.

As the objective is to keep this as a short technical note we refer to the very nice introduction of [2] for a discussion of the background of the problem, its connection to results in various branches of circuit complexity and structural complexity theory, as well as the historical developments leading up the current results.

Let us, however, stress the fact that the current work builds very much on the work of [2] and we use the same framework focused on restrictions and their generalization in the form of projections.

2 An outline of the paper

The function F_d we prove is difficult to compute on average is more or less the first function that comes to mind and slight variants of this function are used in earlier paper proving hierarchy results. It is computed by a read-once formula which is a tree of depth d and governed by a parameter $m \approx \log n/2d$. We have alternating levels of and-gates and or-gates, the top fanin is $\Theta(2^{2m})$, the bottom fanin is $2m$ while the fanin at other levels is $\Theta(m2^{2m})$. A little care is needed to make it balanced but this is straightforward.

We are given a depth $d-1$ circuit C that we wish to prove does not even approximate F_d. The early papers [3, 4, 1] picked a random restriction (giving values to most variables) keeping a few variables non-assigned and made sure that F_d turned into F_{d-1} while the depth of the circuit C could be decreased. A new element was introduced by [2] by allowing projections (identifying several different old variables with one new variable) but this paper also relied on induction.

We follow almost the same approach but there is a subtle difference. After we have done a random restriction and projection, F_d does not turn exactly into F_{d-1} but to a similar looking function and previous papers uses a clean-up stage to turn it exactly into F_{d-1} in order to apply the induction hypothesis. This clean-up is fairly costly and we omit it. Instead, starting with F_d we apply a sequence of $d-1$ restrictions and projections and after the i'th stage F_d is reduced to something similar to, but not exactly equal to F_{d-i}, while C has, with high probability, lost i levels. Apart from this little change the approach of the paper is what can be expected and an outline of the paper is as follows.

Defining the function F_d is done in Section 3. We define our set of restrictions in Section 4. Their basic properties, that they generate uniformly random inputs and that they, with high probability and to a significant extent, preserve F_d is established in Section 5.

One way of thinking about picking the restriction ρ^i is by first picking independent restrictions ρ^{i-1} to each sub-formula of depth $i-1$ and then doing some additional fixing. This is also the way we reason about ρ^i when performing the simplifications, via a switching lemma, of the competing circuit C. The needed switching lemma is found in Section 6. We put all the pieces together proving our main theorem in Section 7.

3 Defining the functions F_d

We have a parameter m guiding our construction. We define F_d to be a tree with alternating layers of and-gates and or-gates. The layer next to the inputs is defined to contain and-gates and thus the output gate is an and-gate if d is odd and otherwise an or-gate.

The intuitive version of the definition is that we want internal and-gates to be 1 with probability 2^{-2m} for random inputs and or-gates to be 0 with the same probability. Furthermore we want the output to be unbiased. This cannot be achieved exactly but by an inductive definition we can come very close. Let us turn to the formal details and give the parameters.

Definition 3.1 *Let $c_0 = \frac{1}{2}$ and for $1 \leq i \leq d-1$ let f_i be the smallest integer such that*

$$(1 - c_{i-1})^{f_i} \leq 2^{-2m},$$

and set $c_i = (1 - c_{i-1})^{f_i}$. Finally set f_d to be the smallest integer such that

$$(1 - c_{d-1})^{f_d} \leq \frac{1}{2}.$$

The function F_d is defined by a read-once formula of alternating levels of and-gates and or-gates. The fan-out at distance i from the inputs is f_i.

It is not difficult to see that $2^{-2m} - 2^{-4m} \leq c_i \leq 2^{-2m}$ for $1 \leq i \leq d-1$ and $f_1 = 2m$, $f_i = 2m \ln 2 \cdot 2^{2m}(1 + O(2^{-m}))$ for $2 \leq i \leq d-1$, while $f_d = \ln 2 \cdot 2^{2m}(1 + O(2^{-m}))$. It follows that the number of inputs of F_d is

$$\prod_{i=1}^{d} f_i = 2^{2(d-1)m} m^{d-1} 2^{O(d)} \tag{1}$$

and we denote this number by n. We note that if $d \leq \frac{\log n}{2 \log \log n}$ then the first factor of (1) is the dominating factor and $m = \frac{\log n}{2d-2}(1 + o(1))$.

It follows by construction that if we feed random independent uniform bits as inputs into the formula defining F_d then an and-gate on level i (which assumes that i is odd) is one with probability c_i while an or-gate is zero with the same probability for even i. It follows that output of F_d is, within an error 2^{-2m}, unbiased. We turn to defining the space of restrictions.

4 The space of restrictions R^i

In this paper we, as was first done in [2], complement restrictions with projections. A classical restriction maps each variable, x_i, to one of the three values 0, 1 and $*$. The two first values indicate that the corresponding constant should be substituted for this variables while the third value says that that the value of x_i remains undetermined.

We combine classical restriction with projections under which groups of variables are identified with the same new variable. This makes further simplifications possible. The mapping of old variables to new variables could be completely arbitrary but to avoid a generality that we do not utilize we define only a special class of projections.

The universe of variables used by our restrictions is given by x_v where v ranges over all nodes in the tree defining F_d. Let V_i be the set of variables x_v where v is at height i, i.e. at distance i from the inputs. Note that the set of original inputs to F_d is exactly given by V_0.

Definition 4.1 *A level i restriction, ρ^i, is a mapping of V_0 into the set $\{0,1\} \cup V_i$. The possible values of $\rho^i(x_w)$ are 0, 1 and x_v where v is the height i ancestor of w.*

The only way we construct a level i restriction in this paper is to first do a level $i-1$ restriction, then apply a classical restriction to the variables in V_{i-1} and finally identify any live variables with its parent. Thus when going from ρ^{i-1} to ρ^i we define a mapping from V_{i-1} to $\{0,1\} \cup V_i$ and ρ^i is the composition of these two mapping. Any input mapped to a constant under ρ^{i-1} is still mapped to the same constant under ρ^i.

A central role in our proof is played by a probability distribution of level i restrictions, R^i, and let us give its basic properties. The restrictions operate independently on each height i sub-formula and let us assume that i is odd and hence the top gate of such a sub-formula is an and-gate. Let F_v the be sub-formula rooted at v, a gate at level i in the formula defining F_d. We have four basic properties of our space of restrictions.

1. With probability $2^{-5m/2}$ all variables of F_v are fixed to constants and $F_v\lceil_{\rho^i} \equiv 1$.
2. With probability $1 - 2^{-m}$ all variables of F_v are fixed to constants and $F_v\lceil_{\rho^i} \equiv 0$.
3. With probability $2^{-m} - 2^{-5m/2}$ we have $F_v\lceil_{\rho^i} \equiv x_v$.
4. If x_v is set to 1 with probability b_i defined as

$$b_i = \frac{c_i - 2^{-5m/2}}{2^{-m} - 2^{-5m/2}} \tag{2}$$

then ρ^i combined with this setting gives a uniformly random input to all variables in F_v.

For future reference we note that $b_i = 2^{-m}(1+O(2^{-m/2}))$. If the gate is an or-gate we reverse the roles of 0 and 1. The spaces of restrictions are constructed recursively for increasing values of i and let us start by formally defining R^1. Let v be a gate at level one and let B_v be the inputs that appear in F_v which is thus a set of size $2m$.

Definition 4.2 *A random restriction $\rho^1 \in R^1$ is constructed independently for each gate, v, on level 1 as follows.*

1. *Pick a uniformly random assignment $\alpha \in \{0,1\}^{2m}$ to all inputs $x_w \in B_v$.*
2. *If $\alpha = 1^{2m}$ then with probability $2^{-m/2}$ set $\rho(x_w) = 1$ for all $x_w \in B_v$, and otherwise proceed as follows. Pick a uniformly random non-empty subset S of B_v and set $\rho(x_w) = *$ for $x_w \in S$ and otherwise set $\rho(x_w) = 1$.*
3. *If $\alpha \neq 1^{2m}$, then with probability $(1-2^{-m})/(1-2^{-2m})$ set $\rho(x_w) = \alpha_w$ for all $x_w \in B_v$ and otherwise set $\rho(x_w) = *$ for all w such that $\alpha_w = 0$ while $\rho(x_w) = 1$ when $\alpha_w = 1$.*
4. *Identify all variables $x_w \in B_v$ such that $\rho(x_w) = *$ with x_v.*

We let ρ^1 denote a typical level 1 restriction after also the projection in the last step has been applied. If some variable x_w is mapped to x_v then we say that x_v is *alive* and also write this as $\rho^1(x_v) = x_v$. Keeping with this convention we also write $\rho^1(x_v) = c$ when F_v is fixed to the constant c. In general we sometimes use $\rho^1(x_v)$ for $F_v\lceil_{\rho^1}$ and remember that this takes values 0, 1 or x_v.

We think of the assignment α as a "tentative" assignment to all variables. We forget some of the values but if we later assign x_v with the correct bias we get the same probability distribution as if we had kept the original α. This assures that such a substitution creates a uniformly random input.

It is not difficult to see that R^1 has the four basic properties we described above but let us still check them in detail.

The probability that F_v is fixed to 1 is $2^{-5m/2}$ as we need to pick $\alpha = 1^{2m}$ and then decide to use this assignment fully under step 2. Similarly the probability that F_v is fixed to 0 is

$$(1-2^{-2m}) \cdot (1-2^{-m})/(1-2^{-2m}) = 1-2^{-m}$$

as this must happen under step 3. This ensures the two first properties. Note that in all other cases we have a non-empty set S such that $\rho(x_w) = x_v$ for all $x_w \in S$ while all other variables of B_v are mapped to 1. This implies that the value at $F_v\lceil_{\rho^1} = x_v$. We now turn to the property that if any live x_v is set to 1 with probability $b_1 = \frac{2^{-2m}-2^{-5m/2}}{2^{-m}-2^{-5m/2}}$ then we have the distribution of a random input.

The probability that a set S is chosen under step 2 is

$$p_2^S = \frac{2^{-2m}(1-2^{-m/2})}{2^{2m}-1}$$

while the probability that the same set is chosen under step 3 is

$$p_3^S = \frac{2^{-2m}(2^{-m} - 2^{-2m})}{(1 - 2^{-2m})} = \frac{(2^{-m} - 2^{-2m})}{(2^{2m} - 1)}.$$

This implies that, conditioned that S is the set of inputs such that $\rho^1(x_w) = x_v$, the probability the probability that it was chosen under step 2 is

$$\frac{p_2^S}{p_2^S + p_3^S} = \frac{2^{-2m} - 2^{-5m/2}}{2^{-m} - 2^{-5m/2}} \tag{3}$$

and this is exactly b_1. This implies that if x_v is set to 1 with probability b_1 then we get the same probability distribution on x as if we had set $x_w = \alpha_w$ immediately. We conclude that we get a uniformly random input to the gate v. We conclude that R^1 has all the desired properties and proceed to the general case.

When picking a restriction in the space R^i we first pick a restriction ρ^{i-1} from R^{i-1} which reduces each sub-formula, F_w of depth $i-1$ to a constant or a variable x_w. This implies that going from ρ^{i-1} to ρ^i is essentially picking the inputs to a single gate and thus quite similar to a restriction from R^1. We again center the construction around a Boolean vector α which plays the role of a set of independent but suitably biased set of values at level $i-1$. The bits of α come in two flavors. Those that are "hard" which should be thought of as already fixed by ρ^{i-1} and hence cannot be changed to x_v and those that are "soft" and can be changed.

Let us assume that i is odd and hence that each gate v on level i is an and-gate. In the case of even i the role of 0 and 1 are reversed in the definition below. As before we let B_v be the set of input gates to v which is now of size f_i.

We first pick an input $\alpha \in \{0,1\}^{f_i}$ where some values are hard while other are soft. For each α_w independently.

1. Make it a *hard zero* with probability $2^{-5m/2}$.
2. Make it a *hard one* with probability $1 - 2^{-m}$.
3. Make it a *soft zero* with probability $c_{i-1} - 2^{-5m/2}$.
4. Make it a *soft one* with probability $2^{-m} - c_{i-1}$.

We note that a coordinate that is not given a hard value is set to a soft zero with probability exactly b_{i-1}.

Let T be the set of inputs that are given soft values. Thus typically T is of size roughly $2^{-m} f_i$ and let f_v be the actual number. Let S be a potential set of soft zeroes. For a non-empty set T, by a "uniformly non-empty subset of T of bias b_{i-1}" we mean that we include each element of T with probability b_{i-1} in S and if S turns out to be empty we try again. We denote this probability distribution by $q_{S,T}$ and it is not difficult to see that

$$q_{S,T} = (1 - (1 - b_{i-1})^{f_v})^{-1} b_i^{|S|} (1 - b_{i-1})^{f_v - |S|}. \tag{4}$$

Note that if T is about its typical size, then the probability (conditioned on picking T) that S is empty is about 2^{-2m} so this conditioning is in general mild. Now we proceed to the determine the value of ρ^i by determining the values of the gates in B_v to be 0, 1 or x_v. As indicated above we never change a hard value while soft values are either made permanent or turned into x_v.

1. If there is at least one hard zero in B_v set $\rho(x_w) = \alpha_w$ for all w.
2. If $|f_v - f_i 2^{-m}| \geq 2^{3m/4}$ then set $\rho(x_w) = \alpha_w$ for all w.

Let us turn to the more interesting part of ρ. Let q_3 and $q_4(f_v)$ be constants, which in rough terms satisfy $q_3 \approx 2^{-m/2}$ and $q_4(f_v) \approx 2^{-m}$, but whose exact values are given during the analysis.

3. If $\alpha = 1^{f_i}$, then with probability q_3 set $\rho(x_w) = 1$ for all w and otherwise proceed as follows. Choose a non-empty subset S of T with bias b_{i-1} and set $\rho(x_w) = *$ for $x_w \in S$ while $\rho(x_w) = 1$ otherwise.
4. Suppose we get a non-empty set S of soft zeroes. Then with probability $1 - q_4(f_v)$ set $\rho(x_w) = \alpha_w$ for all $x_w \in B_v$ and otherwise we set $\rho(x_w) = *$ for $w \in S$ and $\rho(x_w) = 1 = \alpha_w$ otherwise.
5. Identify all live variables in B_v with x_v.

Before proceeding let us observe that if any coordinate α_w is set to a hard value, x_w is always set to this value. This follows as S is a subset of T and hence not given hard ones as values and if a hard zero is assigned, all values of α are used.

The above description tells us how to go from ρ^{i-1} to ρ^i and there are a couple of equivalent ways to view the combination into a full assignment. One way is the following.

1. Pick an assignment α.
2. For hard coordinates w of α pick a restriction $\rho^{i-1} \in R^{i-1}$ conditioned on $\rho^{i-1}(x_w)$ being this constant.
3. For soft coordinates w of α pick a restriction $\rho^{i-1} \in R^{i-1}$ conditioned on $\rho^{i-1}(x_w) = x_w$ and then set $\rho^i(x_w)$ as in the above procedure.

Of course an equivalent way to describe the procedure is to first pick random independent $\rho^{i-1} \in R^{i-1}$ for each depth $i-1$ circuit and then a value of α conditioned on getting hard coordinates with the correct value whenever $\rho^{i-1}(x_w)$ was chosen to be a constant. In more detail this is the following procedure.

1. Pick a random $\rho^{i-1} \in R^{i-1}$. Whenever $\rho^{i-1}(x_w)$ is a constant we fix α_w to be that constant in hard way.
2. For any w such that α_w is not set in step 1 pick it to be a soft value with bias b_{i-1}.

3. Fix the values of some x_w to constants based on cases 1 and 2. This is done by a traditional restriction taking values $0, 1$ and $*$ and we denote this restriction by ρ_1.

4. Fix the values of some x_w for $w \in V_{i-1}$ based on cases 3 and 4. This as a traditional restriction that we denote by ρ.

5. For all $x_w \in B_v$ with $\rho(x_w) = *$ set $\rho^i(x_w) = x_v$. We call this projection π.

We say that ρ^{i-1}, ρ_1, ρ and π are the *components* of ρ^i. The most interesting part when going from ρ^{i-1} to ρ^i turns out to be the third step ρ. For a function f we let $f\lceil_\rho$ denote the function after this step. We let $f\lceil_{\rho+\pi}$ denote the function also after the projection has been made. We have that $f\lceil_\rho$ is a function of x_w where $w \in V_{i-1}$ while $f\lceil_{\rho+\pi}$ is a function of x_v where $v \in V_i$.

As an example suppose that $f = x_{w_1} \vee \bar{x}_{w_2}$ where w_1 and w_2 are two nodes in the same depth i sub-formula. Suppose furthermore that ρ does not fix either of these variables. In this situation $f\lceil_\rho$ is the same function as f while $f\lceil_{\rho+\pi}$ is identically true.

5 Simple properties of R^i

The construction of the space of restrictions, R^i has been carefully crafted to, more or less by definition, satisfy the four basic properties and we establish these as a sequence of lemmas.

Lemma 5.1 *For any node v on level i in F_d such that x_v is alive after ρ^i we have $F_v\lceil_{\rho^i} = x_v$. Furthermore if $F_v\lceil_{\rho^i}$ is a constant then ρ^i assigns constants to all variables in F_v.*

Proof: Going over the construction line by line it is not difficult to see that this is true. ∎

The second lemma says that a sub-formula is reduced to 0 with the correct probability.

Lemma 5.2 *We can determine a value of $q_3 = 2^{-m/2}(1 + o(m))$ such that $Pr[F_v\lceil_{\rho^i} = 1] = 2^{-5m/2}$.*

Proof: Let p_2 be the probability that that case 2 happens. By standard Chernoff bounds we have $p_2 = \exp(-\Omega(2^{m/2}/m))$. Let $p_{2,1}$ be the probability that the value of F_v is fixed to 1 under case 2.

Now let p_3 be the event that we are in case 3. As the probability that $\alpha = 1^{f_i}$ is c_i we have that $p_3 = c_i - p_{2,1}$ and thus $p_3 = 2^{-2m}(1 + o(m))$. Fix the value of q_3 to be $(2^{-5m/2} - p_{2,1})/p_3$ and note that $q_3 = 2^{-m/2}(1 + o(m))$ as promised. The probability of fixing F_v to the value 1 is $p_{2,1} + p_3 q_3$ and this, by the choice of q_3, equals $2^{-5m/2}$. ∎

We next determine a suitable value for $q_4(f_v)$.

Lemma 5.3 *We can determine a value of $q_4(f_v) = 2^{-m}(1 + o(m))$ such that setting $x_v = 1$ with probability b_i gives the same distribution as setting $x_w = \alpha_w$ for all w.*

Proof: We prove that for any S and T, conditioned on T being the non-hard ones and S being the set of variables set to x_v, the probability that this happened in case 3 is b_i while the probability that this happened in case 4 is $1 - b_i$. As we in case 3 change the values of the variables in S from 1 to x_v and in case 4 from 0 to x_v this is sufficient to prove the lemma. From now on we condition on a particular set T of size f_v being chosen and no hard zero being picked.

The conditional probability of ending up with a specific set S and being in case 3 is

$$p_{S,T}^{4,3} = (1 - b_{i-1})^{f_v} q_{S,T},$$

where $q_{S,T}$ is the conditional probability as in (4). The probability of getting the same sets in case 4 is

$$p_{S,T}^{4,4} = (1 - (1 - b_{i-1})^{f_v}) q_{S,T} q_4(f_v).$$

We now set

$$q_4(f_v) = \frac{(1 - b_{i-1})^{f_v}(1 - b_i)}{(1 - (1 - b_{i-1})^{f_v}) b_i} \tag{5}$$

with the result that

$$\frac{p_{S,T}^{4,4}}{p_{S,T}^{4,3}} = \frac{1 - b_i}{b_i}. \tag{6}$$

Since we did not fix the values of all variables in case 2 we know that $f_v = f_i 2^{-m}(1 + o(m))$ and hence $(1 - b_{i-1})^{f_v} = 2^{-2m}(1 + o(m))$. Furthermore as $b_i = 2^{-m}(1+o(m))$ it is possible to satisfy (5) with $q_4(f_v) = 2^{-m}(1+o(m))$. ∎

From now on we assume that we use the values of q_3 and $q_4(f_v)$ determined by Lemma 5.2 and Lemma 5.3. The next lemma is, more or less, an immediate consequence of Lemma 5.3.

Lemma 5.4 *Let v be a node on level i of the formula for F_d. Then picking a random $\rho^i \in R^i$ and then setting x_v with bias b_i gives the uniform distribution on inputs to the formula F_v.*

For completeness let us point out that by "bias b_i" we here mean that x_v is more likely to be 0 if i is odd and more likely to be 1 if i is even.

Proof: We proceed by induction over i and for $i = 1$ we already established the lemma in the discussion after the definition of R^1.

Lemma 5.3 tells us that picking a x_v with bias b_i is the same as setting the soft values according to α. This, in its turn, is the same as picking independent restrictions from ρ^{i-1} for each sub-formula F_w and setting any live x_w with bias b_{i-1}. By induction this results in the uniform distribution. ∎

Let us also verify the last basic property of R^i.

Lemma 5.5 *We have* $Pr[F_v\lceil_{\rho^i}= 0] = 1 - 2^{-m}$.

Proof: This could be done by a tedious calculation, but in fact it can be seen by a high level argument. The restriction ρ^i can reduce F_v to 0, 1 or x_v. Lemma 5.2 says that second value is taken with the correct probability and Lemma 5.4 says that if x_v is set to 1 with bias b_i then we get a uniformly random input and hence the output of F_v is one with probability c_i. This implies that

$$2^{-5m/2} + b_i Pr[F_v\lceil_{\rho^i} \equiv x_v] = c_i$$

and hence, by the definition of b_i, we conclude that $Pr[F_v\lceil_{\rho^i} \equiv x_v] = 2^{-m} - 2^{-5m/2}$ and as the probabilities of obtaining the three possible values for $F_v\lceil_{\rho^i}$ sum to one, the lemma follows. ∎

The most interesting property of our restrictions is that we can prove a switching lemma and we proceed with this step.

6 The switching lemmas

To establish a general hierarchy result (in particular distinguishing depth d and $d-2$) it is sufficient to prove a switching lemma for R^i for $i \geq 2$ and in view of this we prove this lemma first. To get a tight result we later prove a modified lemma for R^1.

As discussed in Section 4, a restriction $\rho^i \in R^i$ is chosen by first picking $\rho^{i-1} \in R^{i-1}$, followed by ρ_1, ρ and finally making a projection π. In this section we assume any fixed values of of ρ^{i-1} and ρ_1 and consider the effect of ρ. The fact that the distribution of ρ is dependent on the actual values of ρ^{i-1} and ρ_1 is left implicit.

A set $\mathcal{F}$ of restrictions is said to be "downward closed" if changing the value of ρ on some input from the value $*$ to a constant cannot make it leave the set. Let us write this formally.

Definition 6.1 *A set $\mathcal{F}$ of restrictions is* downward closed *if when $\rho \in \mathcal{F}$ and $\rho'(x_w) = \rho(x_w)$ for $w \neq w_0$ and $\rho(x_{w_0}) = *$ then $\rho' \in \mathcal{F}$.*

We can now formulate the main lemma.

Lemma 6.2 *Let $\rho^i \in R^i$ be a random restriction with components ρ^{i-1}, ρ_1, ρ and π and f be an arbitrary function. Suppose $g = f\lceil_{\rho^{i-1}}$ is computed by a depth-2 circuit of bottom fanin $t \leq 2^m/8$. Let $\mathcal{F}$ be a downward closed set of restrictions and let $depth(g\lceil_{\rho^i})$ be the minimal depth of the decision tree computing $g\lceil_{\rho^i}$. Then, for sufficiently large m,*

$$Pr[depth(g\lceil_{\rho^i}) \geq s \mid \rho \in \mathcal{F}] \leq D^s,$$

where $D = t2^{3-m/2}$.

Proof: By symmetry we may assume that i is odd. By possibly looking at the negation of g (which has the same depth decision tree as g) we can assume that g is a CNF, and after ρ_1 has been applied it can be written as

$$g\lceil_{\rho_1} = \wedge_{i=1}^{\ell} C_i,$$

where each C_i is a disjunction of at most t literals. The proof proceeds by induction over ℓ and the base case is when $\ell = 0$ in which case $g\lceil_{\rho^i}$ is always computable by a decision tree of depth 0.

We divide the analysis into two cases depending on whether C_1 is forced to one or not. We can bound the probability of the lemma as the maximum of

$$Pr[\text{depth}(g\lceil_{\rho^i}) \geq s \mid \rho \in \mathcal{F} \wedge C_1\lceil_\rho \equiv 1],$$

and

$$Pr[\text{depth}(g\lceil_{\rho^i}) \geq s \mid \rho \in \mathcal{F} \wedge C_1\lceil_\rho \not\equiv 1]. \tag{7}$$

The first term is taken care of by induction applied to g without its first conjunction (and thus having size at most $\ell - 1$) and using that the conditioning in this case is a new downward closed set. We need to consider the second term (7).

To avoid that $g\lceil_{\rho^i} \equiv 0$ there must be some non-empty set, Y of variables appearing in C_1 which are given the value $*$ by ρ. Let a set B_v of variables be called a "block" and suppose the variables in C_1 come from t_1 different blocks. Say that a block is "undetermined" if it contains a variable given the value $*$ by ρ. Let Z be the set of undetermined blocks and let us assume it is of size r. Let us introduce the notation $undet(Z)$ to denote the event that all blocks in Z are undetermined and $det(C_1/Z)$ to say that all variables in C_1 outside Z are fixed to non-$*$ values by ρ.

We start constructing a decision tree for $g\lceil_{\rho^i}$ by querying the new variables corresponding to Z. Let τ be an assignment to these variables. We can now bound (7) as

$$\sum_{\tau,Z} Pr[\text{depth}(g\lceil_{\tau\rho^i}) \geq s - r \wedge undet(Z) \wedge det(C_1/Z) \mid \rho \in \mathcal{F} \wedge C_1\lceil_\rho \not\equiv 1], \tag{8}$$

where r is the size of Z which is non-empty. We will use the estimate

$$\begin{aligned} Pr[\text{depth}(g\lceil_{\tau\rho^i}) \geq s - r \mid undet(Z) \wedge det(C_1/Z) \wedge \rho \in \mathcal{F} \wedge C_1\lceil_\rho \not\equiv 1] \times \\ Pr[undet(Z) \mid \rho \in \mathcal{F} \wedge C_1\lceil_\rho \not\equiv 1] \end{aligned} \tag{9}$$

for each term in (8) and hence a key lemma is the following.

Lemma 6.3 *If Z is a set of set r blocks appearing in C_1 and ρ a random restriction appearing in the construction of R^i for $i \geq 2$, then, for sufficiently large m,*

$$Pr[undet(Z) \mid \rho \in \mathcal{F} \wedge C_1\lceil_\rho \not\equiv 1] \leq 2^{r(1-m/2)}.$$

Proof: The crux of the proof is to, given a restriction ρ that contributes to the probability in question, create a restriction ρ' that also satisfies the conditioning but fixes all variables in the blocks of Z. We describe how to do this for $r = 1$ but the general case follows immediately as we can do the changes independently on each block. Thus let us assume that Z is the single block B_v and fix a restriction ρ that contributes to the event of the lemma. Let P be the set of variables of B_v that appears positively in C_1 and N the set of variables that appear negatively.

We can assume that we have no hard zero in B_v and the number of non-hard ones in B_v is close to $f_i 2^{-m}$ as otherwise already ρ_1 would have fixed all variables in B_v to constants.

Clearly for ρ we must have $\rho(x_v) = *$. For variables $x_w \in P$ we must have $\rho(x_w) = *$ while for variables in N we have either $\rho(x_w) = *$ or $\rho(x_w) = 1$.

We now define a companion restriction $\rho' = H(\rho)$. If ρ maps some variable outside N to $*$ (and in particular if P is non-empty) we set $\rho'(x_v) = 0$ and otherwise $\rho'(x_v) = 1$. For a $x_w \in P$ we set $\rho'(x_w) = 0$ while for $x_w \in N$ we set $\rho'(x_w) = 1$, independently of the value of $\rho(x_w)$. Outside C_1 but in B_v we set $\rho'(x_w) = 1$ if $\rho(x_w) = 1$ and $\rho'(x_w) = \rho'(x_v)$ otherwise. Outside B_v, ρ and ρ' agree. First observe that ρ' satisfies the conditioning. We only have $\rho(x_w) \neq \rho'(x_w)$ when $\rho(x_w) = *$ and by the definition of P and N we are careful not to satisfy C_1.

The mapping H is many-to-one as given ρ' we do not know the values of $\rho(x_w)$ when $x_w \in N$ (but we do for all other variables in B_v).

First note that

$$\frac{Pr(\rho)}{Pr(\rho')} = \frac{Pr(\rho_v)}{Pr(\rho'_v)}$$

where ρ_v is only the behavior of ρ on B_v and similarly for ρ'. This is true as ρ and ρ' take the same values outside B_v and the restrictions are picked independently on each B_v.

Assume first that $\rho'(x_v) = 0$. In this situation ρ could have been picked under case 3 or case 4 while ρ' can only have been produced under case 4. We know, by (6), that each ρ is about a factor 2^m more likely to have been produced under case 4 than under case 3 so let us ignore case 3, introducing a small error factor $(1 + O(2^{-m}))$ that we temporarily suppress.

Let N_1 be subset of N that was actually given the value $*$ by ρ. If N_1 is empty then $Pr(\rho_v) = \frac{q_4(f_v)}{1-q_4(f_v)} Pr(\rho'_v)$ and in general we pick up an extra factor $b_{i-1}^{|N_1|}(1 - b_{i-1})^{-|N_1|}$. As

$$\sum_{N_1 \subseteq N} b_{i-1}^{|N_1|}(1 - b_{i-1})^{-|N_1|} = (1 + \frac{b_{i-1}}{1 - b_{i-1}})^{|N|}$$

we get

$$\sum_{H(\rho)=\rho'} Pr(\rho) \leq \left(1 + \frac{b_{i-1}}{1 - b_{i-1}}\right)^{|N|} \frac{q_4(f_v)}{1 - q_4(f_v)} Pr[\rho'] \leq 2^{1-m} Pr[\rho'], \qquad (10)$$

for sufficiently large m. This follows as $|N| \leq 2^m/8$, $b_i = (1+o(m))2^{-m}$ and $q_4 = (1+o(m))2^{-m}$.

If, $\rho'(x_v) = 1$ the situation is similar except that ρ' is produced under case 3 and thus we pick up a factor q_3 instead of $1 - q_4(f_v)$. We get in this case

$$\sum_{H(\rho)=\rho'} Pr(\rho) \leq \left(1 + \frac{b_{i-1}}{1-b_{i-1}}\right)^{|N|} \frac{q_4(f_v)}{q_3} Pr[\rho'] \leq 2^{1-m/2} Pr[\rho'], \tag{11}$$

again for sufficiently large m. The fact that we ignored restrictions ρ produced under case 3 gives an additional factor $(1 + O(2^{-m}))$ in the above estimates and thus the calculations remain valid, possibly by making m slightly larger, to make sure that the "sufficiently large m" statements are true.

The case of general r follows from the fact that we do the modifications on all blocks of Z independently. ∎

Remark 1 *The careful reader might have noticed that in the case with the $\rho'(x_v) = 1$ then we can conclude that N_1 is non-empty giving a slightly better estimate especially in the case when t is small. This observation can probably be used to get a slightly better constant in the main theorem, but to keep the argument simple we ignore this point. We return to main argument.*

We now estimate

$$Pr[\text{depth}(g\lceil_{\tau\rho^i}) \geq s - r \mid undet(Z) \wedge det(C_1/Z) = 0 \wedge \rho \in \mathcal{F} \wedge C_1\lceil_\rho \not\equiv 1], \tag{12}$$

by induction. We need to check that the conditioning defines a downward closed set. This is not complicated but let us spell out some details. Fix any behavior of ρ inside the blocks of Z and satisfying the conditioning. As $g\lceil_{\rho^i\tau}$ does not depend on the variables corresponding to Z the event in (12) depends only the values of ρ outside Z. Changing ρ from $*$ to a constant value for any variable outside Z cannot violate any of the conditions in the conditioning and hence we have a downward closed set when considering ρ as a restriction outside Z. We conclude that the probability of the event in (12) is, by induction, bounded by D^{s-r}.

Our goal is to estimate the sum (8), using the bound (9) for each term, Lemma 6.3 and the inductive case. If C_1 intersects t_1 different blocks (where of course $t_1 \leq t$) then, using the fact that we have at most 2^r (remember that r is the number of blocks of Z) different τ, we get the total estimate

$$\sum_{Z \neq \emptyset} 2^r 2^{r(1-m/2)} D^{s-r} = D^s \left((1 + D^{-1}2^{2-m/2})^{t_0} - 1\right) \leq D^s \left((1 + \frac{1}{2t})^{t_0} - 1\right) \leq D^s$$

and we are done. ∎

Lemma 6.2 is sufficient to prove a fairly tight hierarchy theorem. To prove a tight variant we need also to see how R^1 simplifies circuits.

Lemma 6.4 *Let g be computed by a depth-2 circuit of bottom fanin $t \leq m/4$. Let $\mathcal{F}$ be a downward closed set of restrictions and ρ^1 a random restriction with the distribution R^1. Let $depth(g\lceil_{\rho^1})$ be the minimal depth of the decision tree computing $g\lceil_{\rho^1}$. Then, for sufficiently large m,*

$$Pr[depth(g\lceil_{\rho^1}) \geq s \mid \rho \in \mathcal{F}] \leq D^s,$$

where $D = t2^{3+t-m/2}$.

Proof: The proof of this lemma is almost identical to the proof of Lemma 6.2 and let us only discuss the differences. Lemma 6.3 is replaced by the following.

Lemma 6.5 *If Z is a set of set r blocks appearing in C_1 and ρ a random restriction appearing in the construction of R^1, then, for sufficiently large m,*

$$Pr[undet(Z) \mid \rho \in \mathcal{F} \wedge C_1\lceil_\rho \not\equiv 1] \leq 2^{r(t+1-m/2)}.$$

Proof: The proof is almost the same as the proof of Lemma 6.3. The reason for the loss in parameters is that the factor

$$\left(1 + \frac{b_{i-1}}{1 - b_{i-1}}\right)^{|N|}$$

that used to be bounded by a constant strictly less than two can now be as large as 2^t. ∎

The rest of the proof of how Lemma 6.4 follows from Lemma 6.5 is identical with how Lemma 6.2 followed from Lemma 6.3 with the obvious change in the final calculation. ∎

7 The proof of the main theorem

We now proceed to prove Theorem 1.1. In fact we are going to prove the following, slightly stronger, theorem.

Theorem 7.1 *Let C be a circuit depth d with bottom fanin at most $m/4$ and which is of size S then, for sufficiently large m,*

$$Pr[F_d(x) = C(x)] \leq \frac{1}{2} + O(2^{-m/4}) + S2^{-2^{m/2-4}}.$$

It is not difficult to see that this theorem implies Theorem 1.1 as a depth $d-1$ circuit can be seen as a depth d circuit with bottom fanin one and that $m = \frac{\log n}{2d-2}(1 + o(1))$. We turn to proving Theorem 7.1.

Proof: Let us apply a random restriction $\rho^{d-1} \in R^{d-1}$ to both F_d and C. Let us assume that i is odd and hence the output gate of F_d is an and-gate. The case of even i is completely analogous. By Lemma 5.4 we have

$$Pr[F_d(x) = C(x)] = Pr[F_d\lceil_{\rho^{d-1}}(x) = C\lceil_{\rho^{d-1}}(x)]$$

where the latter probability is over a random ρ^{d-1} and random assignment to the live variables in V_{d-1} where each variable is given the value 1 with probability $1 - b_{d-1}$. Let us first see how ρ^{d-1} affects F_d.

We have the output gate, v, of fanin f_d. With probability $O(2^{-m/2})$ some input gate is forced to 0 by ρ^{d-1}. Suppose that this does not happen and let h_1 be the number of input gates to v that are not fixed to one. With probability $1 - exp(-\Omega(2^{m/2}))$ we have $|h_1 - f_d 2^{-m}| \leq 2^{3m/4}$. Thus we conclude that, with probability $1 - O(2^{-m/2})$, F_d has been reduced to an and-gate of fanin $\ln 2 \cdot 2^m(1 + O(2^{-m/4}))$.

Now let us see how ρ^{d-1} affects C. We to prove by induction that ρ^i, with high probability, reduces the depth of C by i. Let us assume that C has S_i gates at distance i from the inputs.

Consider any gate in C at distance two from the inputs and suppose it is an or of and-gates, the other case being similar. By Lemma 6.4, for sufficiently large m, after ρ^1 has been applied, except with probability $2^{-2^{m/2-4}}$ this sub-circuit can be computed by a decision tree of depth of depth at most $2^{m/2-4}$. This implies that it can we written as an and of or-gates of fan-out at most $2^{m/2-4}$. We conclude that except with probability $S_2 2^{-2^{m/2-4}}$, by collapsing two adjacent levels of and-gates, $C\lceil_{\rho^1}$ can be computed by a depth $d-1$ circuit with bottom fanin at most $2^{m/2-4}$ where each gate at distance at least two from the inputs corresponds to a gate at distance at least three in the original circuit.

Applying Lemma 6.2 for $i = 2, 3 \ldots d-2$ in a similar way we conclude that except with probability $\sum_{i=3}^{d-2} S_i 2^{-2^{m/2-4}}$, $C\lceil_{\rho^{d-2}}$ can be computed by a depth 2 circuit of bottom fanin $2^{m/2-4}$. A final application of Lemma 6.2 says that except with an additional failure probability $2^{-2^{m/2-4}}$, $C\lceil_{\rho^{d-1}}$ can be computed by a decision tree of depth $2^{m/2-4}$.

By the above reasoning we know that except with probability $O(2^{-m/2}) + S2^{-2^{m/2-4}}$, it is true that $F_d\lceil_{\rho^{d-1}}$ is an and of size $\ln 2 \cdot 2^m(1 + O(2^{-m/4}))$ and $C\lceil_{\rho^{d-1}}$ is computed by a decision tree of depth $2^{m/2-4}$. As the former is equal to 1 with probability $\frac{1}{2}(1 + O(2^{-m/4}))$ and the output of any decision tree of depth s of inputs that are b_{d-1} biased has a sb_{d-1} biased output, we conclude that

$$Pr[F_d\lceil_{\rho^{d-1}}(x) = C\lceil_{\rho^{d-1}}(x)] = \frac{1}{2} + O(2^{-m/4}) + S2^{-2^{m/2-4}}$$

and the proof is complete. ∎

Looking more closely at the proof we can derive an even stronger theorem.

Theorem 7.2 *Suppose d is odd and let C be a circuit depth $d+1$ with output gate that is an or-gate, with bottom fanin at most $m/4$ and of size at most S,* *then, for sufficiently large m,*

$$Pr[F_d(x) = C(x)] \leq \frac{1}{2} + O(2^{-m/4}) + S2^{-2^{m/2-4}}.$$

The same is true for even d if the output gate of C is an and-gate.

Proof: Let us assume that d is odd, the even case being completely analogous. We follow exactly the proof of Theorem 7.1 until the very last step. We can conclude that $C\lceil_{\rho^{d-1}}$, with high probability, is reduced to the disjunction of a set of functions each computable by a decision tree of depth $2^{m/2-4}$. We can convert this to a DNF formula of bottom fanin $2^{m/2-4}$ and we must analyze the probability that such a formula equals an and of size $\ln 2 \cdot 2^m(1 + O(2^{-m/4}))$. We have two cases.

Suppose first that each term in the DNF-formula contains a negated variable. Then $C\lceil_{\rho^{d-1}}$ rejects the all-one input which is chosen with probability $\frac{1}{2} + O(2^{-m/4})$ and as this input is accepted by $F_d\lceil_{\rho^{d-1}}$ we have

$$Pr[F_d\lceil_{\rho^{d-1}}(x) = C\lceil_{\rho^{d-1}}(x)] \leq \frac{1}{2} + O(2^{-m/4}) \tag{13}$$

in this situation (where the probability is only over a random input) and this case follows.

On the other hand if there is a term in $C\lceil_{\rho^{d-1}}$ that only contains positive variables then it (and hence $C\lceil_{\rho^{d-1}}$) is true with probability $1 - O(2^{-m/2})$. As $F_d\lceil_{\rho^{d-1}}$ is close to unbiased, (13) is true also in this case and the theorem follows. ∎

As stated previously we have not done a serious effort to get the best constants in our main theorems. They are, however, not too far from the truth as we may take C to be one input to the output gate of F_d. This is a depth $d - 1$ circuit of sub-linear size that agrees with F_d for a fraction $\frac{1}{2} + \Omega(2^{-2m})$ of the inputs.

8 Some final words

The main difference between the current paper and the early proof of the hierarchy theorem in [1] is the use of projections. The projections serve two purposes. The first is to make sure that once a single $*$ is found in ρ we do not bias any other value of ρ^i to be $*$. This was achieved in [1] by fixing the values of neighboring variables to constants while here we identify all the neighboring variables with the same new variable and hence we only query one variable in the decision tree. We feel that this difference is minor.

The more important difference is that projections enables us to choose a uniformly random input where this seemed difficult to achieve. It is amazing how seemingly simple ideas can take care of problems that, at least initially, looks like fundamental obstacles.

Acknowledgment I am grateful Ben Rossman, Rocco Servedio and Li-Yang Tan for giving me access to an early version of their paper and for fruitful discussions regarding the topic of the current paper. I also thank Per Austrin for coming up with the nice argument that F_d and C are almost un-correlated used in the proof of Theorem 7.2.

References

[1] J. Håstad. Almost optimal lower bounds for small depth circuits. In *Proceedings of the eighteenth annual ACM symposium on Theory of computing*, STOC '86, pages 6–20, New York, NY, USA, 1986. ACM.

[2] Benjamin Rossman, Rocco A. Servedio, and Li-Yang Tan. An average-case depth hierarchy theorem for boolean circuits. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 1030–1048, 2015.

[3] M. Sipser. Borel sets and circuit complexity. In *Proceedings of the fifteenth annual ACM symposium on Theory of computing*, STOC '83, pages 61–69, New York, NY, USA, 1983. ACM.

[4] A. C-C. Yao. Separating the polynomial-time hierarchy by oracles. In *26th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '85, pages 1 –10. IEEE, 1985.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de