Limits on Alternation-Trading Proofs for Time-Space Lower Bounds

Samuel R. Buss*
Department of Mathematics
University of California, San Diego
La Jolla, CA
sbuss@math.ucsd.edu

Ryan Williams†
Computer Science Department
Stanford University
Stanford, CA
rrwilliams@stanford.edu

***Abstract*—This paper characterizes alternation trading based proofs that the satisfiability problem is not in the time and space bounded class $\mathrm{DTISP}(n^c, n^\epsilon)$, for various values $c < 2$ and $\epsilon < 1$. We characterize exactly what can be proved for $\epsilon \in o(1)$ with currently known methods, and prove the conjecture of Williams that the best known lower bound exponent $c = 2\cos(\pi/7)$ is optimal for alternation trading proofs. For general time-space tradeoff lower bounds on satisfiability, we give a theoretical and computational analysis of the alternation trading proofs for $0 < \epsilon < 1$, again proving time lower bounds for various values of ϵ which are optimal for the alternation trading proof paradigm.**

I. Introduction

How powerful are the methods we have for proving computational lower bounds? What prevents us from proving major complexity class separations? Over the years, there have been several formalizations of these questions that led to new insights into complexity itself. Relativization shows the limits of proving lower bounds via naive diagonalization [2], [11], natural proofs show the limits of combinatorial arguments in circuit complexity [13], and algebrization shows the limits of low-degree polynomial techniques [1], [8]. Understanding the power of existing methods is a fundamental issue in complexity. By determining the limits of what is provable with known methods, we can discover how to improve upon their weaknesses.

In this paper, we perform a fine-grained study of a proof method dubbed "alternation-trading", which has been applied to prove many lower bounds via indirect diagonalization arguments. While our approach is not as general as other barrier results, the significant advantage is that we can actually prove "tight" results on what lower bounds can be established. More precisely, our framework produces results of the form: *there is an alternation trading proof that at least n^c resources are necessary to solve problem X, and no alternation trading proof can show that $n^{c+\epsilon}$ resources are necessary to solve problem X, for every $\epsilon > 0$.* Given the scope of alternation trading proofs, we believe that the methods developed here will have further applications to establishing barriers and mining insights into complexity lower bounds.

We focus in this paper on lower bounds for simulating nondeterminism with time- and space-bounded deterministic algorithms, concentrating on the satisfiability problem SAT. However, the known alternation-trading methods for proving these lower bounds also imply analogous time-space lower bounds for many other NP-complete problems (see [17]).

Let $\mathrm{DTISP}(n^c, n^\epsilon)$ denote the class of languages recognizable by deterministic algorithms that run in time $n^{c+o(1)}$ with space bounded by $n^{\epsilon+o(1)}$, where $1 \leq c$ and $0 \leq \epsilon \leq c$. A series of results, see [9], [5], [10], [7], [6], [14], [4], [18], [19], [20], have established better and better non-trivial constant lower bounds on the values c and ϵ for which $\mathrm{SAT} \in \mathrm{DTISP}(n^c, n^\epsilon)$, and similar results for other hard problems. Surveys of these and other results are given by Van Melkebeek [15], [16] but, loosely speaking, all of these lower bounds have been obtained by combining a "speedup" technique of Nepomnjascii [12] with an assumption such as $\mathrm{SAT} \in \mathrm{DTISP}(n^c, n^\epsilon)$ in order to derive a contradiction. For some time, it was a folklore conjecture that these alternation trading proofs could potentially establish that $\mathrm{SAT} \notin \mathrm{DTISP}(n^{2-o(1)}, n^{o(1)})$.

Williams [19], [20] gave a formal definition of these proof methods, which he called "alternation trading proofs", and gave improved time-space lower bounds for SAT. In [20] he designed computer programs that searched for good alternation trading proofs. He conjectured that the proofs found were optimal for alternation trading proofs. The conjecture was somewhat provocative, because the computer searches were far from exhaustive, and the proofs found only established $\mathrm{SAT} \notin \mathrm{DTISP}(n^{2\cos(\pi/7)-o(1)}, n^{o(1)})$ where $2\cos(\pi/7) \approx 1.8019$. This matched the previous lower bound [19] that was felt to be suboptimal.

A. Main Results

Our first main result is a proof of Williams' conjecture: when $\epsilon = 0$, the lower bounds obtained by Williams [19], [20] are in fact *optimal* within the framework of alternation trading proofs. In the course of the proof, we give some surprising simplifications of alternation trading proofs,

* Supported in part by NSF grants DMS-0700533 and DMS-1101228.
† Supported in part by the David Morgenthaler II Faculty Fellowship.

characterizing the possible alternation trading proofs with a device we call "achievable pairs".

Our second main result is to establish new simultaneous time and space lower bounds on deterministic algorithms using alternation trading proofs, along with computer-aided proofs that these lower bounds are optimal in the alternation trading framework. Prior work on time-space tradeoffs includes [14], [4], [6], [19], [20]. In particular, [6] showed that if SAT $\in$ DTISP(n^c, n^ϵ) then $c + \epsilon \geq 1.573$, and [19], [20] improved this to $c + \epsilon \geq 2\cos(\pi/7)$, giving better bounds for specific numeric values of c and ϵ.

The present paper substantially generalizes the prior results by giving a new characterization of arbitrary alternation trading proofs in terms of "achievable triples" which account for arbitrary space bounds of the form n^ϵ. We present a new type of computer-based search for alternation trading proofs via achievable triples, aided by theorems describing how the search space can be pruned. As a consequence, we discover better time-space tradeoffs than those found (and conjectured to be optimal) by [20]. Our computer-based proofs always succeed in establishing either the existence or non-existence of alternation trading proofs for specified time and space bounds. Therefore, our new time-space bounds are in fact the best attainable with presently-known proof methods.

The lower bounds in this paper are all stated for a single problem, SAT. As remarked above, they also apply to many other NP-complete problems. In addition, by [19], our lower bounds also apply to the problem MOD$_m$-SAT of counting the number of satisfying assignments modulo m, where either m is not a prime power or m is prime, with the possible exception of a single prime.

II. Detailed Overview

We now present a more detailed overview of the results. Let DTS(n^c) be the class DTISP(n^c, n^0), i.e. the set of languages accepted by a deterministic random-access machine with runtime $n^{c+o(1)}$ using space $n^{o(1)}$. Refining the proof methods of earlier work, Williams [19], [20] proved that SAT $\notin$ DTS(n^c) for $c < 2\cos(\pi/2)$. He used bounded quantifier notation of the forms "$(\forall n^a)^b$" and "$(\exists n^a)^b$" for constants $a, b \geq 0$, to denote a computation that makes $n^{a+o(1)}$ universal (resp., existential) choices, and then (deterministically) keeps $n^{b+o(1)}$ bits of information. Thus, for instance, $(\exists n^2)^1$DTS(n^3) denotes the class of languages accepted by an algorithm that guesses $n^{2+o(1)}$ bits existentially, deterministically selects $n^{1+o(1)}$ bits to keep in memory (in time $n^{1+o(1)}$), and then runs deterministically in time $n^{3+o(1)}$, using $n^{o(1)}$ workspace in addition to the $n^{1+o(1)}$ bits that were kept as input to the final stage.

Williams [20] presented a general framework for establishing lower bounds via alternation-trading, based on a formal proof system of inference rules that act on bounded quantifier notations for complexity classes. One kind of inference rules are "speedup" rules, which use Nepomnjascii's method of decreasing runtime at the cost of adding alternation(s) to the computation. The second kind of inference rules, called "slowdown" inferences, use the assumption that NTIME$(n) \subseteq$ DTS(n^c) (which follows from SAT $\in$ DTS(n^c)) to remove alternations at the cost of slower runtime. Using the nondeterministic time hierarchy theorem, an alternation trading proof yields a contradiction by providing a proof that DTS$(n^a) \subseteq$ DTS$(n^{a'})$ for constants $a > a' > 0$.

In this framework, binary strings, called "proof annotations", represent patterns of speedup and slowdown inferences in an alternation trading proof, with "**1**" representing a speedup and "**0**" a slowdown. For instance, the annotation **100** represents the sequence *speedup-slowdown-slowdown*; that is, a proof with the rough form:

$$\begin{aligned} &\text{DTS}[n^a] \\ &\quad\subseteq\quad (\exists\ n^{x_1})(\forall\ n^{o(1)})\text{DTS}[n^{a-x_1}] \\ &\quad\subseteq\quad (\exists\ n^{x_1})\text{DTS}[n^{c(a-x_1)}] \\ &\quad\subseteq\quad \text{DTS}[n^{\max\{cx_1, c^2(a-x_1)\}}]. \end{aligned}$$

Setting $a > \max\{cx_1, c^2(a - x_1)\}$ yields a contradiction. Lipton and Viglas [10] proved that in this situation, the optimal setting of parameters yields $c^2 < 2$.

Let $X_0 := \mathbf{(10)^*}$ represent an arbitrary number of speedup-slowdown inferences. Then let X_{i+1} be the annotation $\mathbf{1}X_i\mathbf{0}X_0$. Williams [19] proved these patterns of inferences, as i increases, give contradictions for c arbitrarily close to $2\cos(\pi/7)$, and conjectured in [20] they are the best possible inference patterns that can derived with the formalized speedup and slowdown rules.

We prove these conjectures as Theorem 1. The inference rules R0-R2 are defined below in Section III.

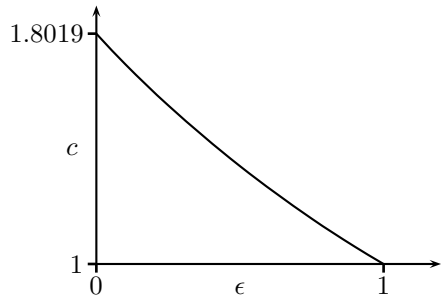
Theorem 1: The inference rules R0–R2 can be used to derive a contradiction to SAT $\in$ DTS(n^c) only for $c < 2\cos(\pi/7)$.

The proof of Theorem 1 is based on a new analysis of what is possible with alternation trading proofs. The central innovation is the concept of "c-achievable pairs" which describe inferences that can be *approximated* by alternation trading proofs that SAT $\notin$ DTS(n^c). Informally, a single c-achievable pair can capture infinite sequences of proof annotations, defined inductively. (Such an infinite sequence of proofs may *converge* to a certain lower bound exponent, but no single proof annotation ever achieves it – hence we speak of alternation trading proofs "approximating" an inference by a c-achievable pair.) Understanding the power of c-achievable pairs turns out to be sufficient for understanding the limits of alternation-trading proofs. We give methods for generating c-achievable pairs, and prove that these pairs exactly characterize the refutations that can be approximated by alternation trading proofs.

The full version of the paper also considers lower bounds on DTISP(n^c, n^ϵ) algorithms for satisfiability, where $\epsilon > 0$

can vary. For these algorithms, we use "(c, ϵ)-achievable triples" that exactly characterize the alternation trading derivations in the DTISP setting. Unlike the $\epsilon = 0$ case, we are unable to give a closed form formula for alternation trading proofs that satisfiability is not in DTISP(n^c, n^ϵ). Instead, we use new computer-based searches for (c, ϵ)-achievable triples that prove the existence of alternation trading refutations. This potentially requires considering infinitely many triples, so to prune the search space, we develop a notion of when two triples together "dual-subsume" a third triple, as well as a related notion of "multisubsumption". These structure theorems allow the computer-based searches to search for quite long proofs. In fact, the computer-based search has always succeeded in finding an alternation trading refutation, or in completely exhausting the search space, proving that there can be no such refutation.

Theorem 2: For all pairs (ϵ, c) in the figure below, SAT $\notin$ DTISP(n^c, n^ϵ), and there is *no* alternation-trading proof of SAT $\notin$ DTISP$(n^{c+0.00001}, n^\epsilon)$.



ϵ	c
0.001	1.80083
0.01	1.79092
0.1	1.69618
0.25	1.55242
0.5	1.34070
0.75	1.15765
0.9	1.06011
0.99	1.00583
0.999	1.00058

Figure 1. Maximum c to five digits of accuracy, as a function of ϵ, for which alternation trading proofs can show SAT $\notin$ DTISP(n^c, n^ϵ).

The outline of this conference version is as follows:

- Section III introduces the speedup and slowdown rules, and alternation trading proofs for SAT $\notin$ DTS(n^c). We give simplified notions of alternation trading proofs, called "h-derivations" and "reduced" derivations, and simplifications of the speedup and slowdown rules.
- Section IV introduces approximate inferences, and the notion of a "c-achievable pair", which informally represent infinite (inductively defined) sequences of alternation proofs. We give intuition for c-achievable pairs and recall the $2\cos(\pi/7)$ lower bound.
- Section V puts limits on what kinds of pairs are c-achievable. Section VI establishes a certain normal form for c-achievable pairs, and completes the proof of Theorem 1.

We review notation and results from earlier work as needed; however, we presume a certain level of familiarity with prior work such as that of Williams [20]. Omitted proofs can be found in the full version of the paper [3].

III. Rules of inference for DTS

A. Basic rules of inference for DTS bounds

Fix, henceforth, a value $c > 1$. The goal is to prove a contradiction from the assumption SAT $\in$ DTS(n^c), thereby of course proving that SAT $\notin$ DTS(n^c). The contradiction is proved by an *alternation trading proof* using the following rules R0–R2. As shown in [20], it suffices to give an alternation trading proof of DTS$(n^a) \subseteq$ DTS(n^b) for some $b < a$. The alternation trading proof is a sequence of containments, starting with the set ${}^1\text{DTS}(n^a)$ for some integer $a > 0$. (The leading superscript "1" indicates the input string has length $1 + o(1)$.)

The following are the original rules of inference used for alternation trading proofs [20]. The ellipses "$\cdots$" indicate an arbitrary (possibly empty) quantifier prefix.

R0: *(Initial speedup)*
${}^1\text{DTS}(n^a) \subseteq {}^1(\exists n^x)^{\max\{x,1\}}(\forall n^0){}^1\text{DTS}(n^{a-x})$,
where $0 < x \leq a$.

R1: *(Speedup)*
$\cdots {}^{b_k}(\forall n^{a_k})^{b_{k+1}}\text{DTS}(n^{a_{k+1}}) \subseteq$
$\cdots {}^{b_k}(\forall n^{\max\{x,a_k\}})^{\max\{x,b_{k+1}\}}(\exists n^0)^{b_{k+1}}\text{DTS}(n^{a_{k+1}-x})$,
where $0 < x \leq a_{k+1}$.

R2: *(Slowdown)*
$\cdots {}^{b_k}(\forall n^{a_k})^{b_{k+1}}\text{DTS}(n^{a_{k+1}}) \subseteq$
$\cdots {}^{b_k}\text{DTS}(n^{\max\{cb_k, ca_k, cb_{k+1}, ca_{k+1}\}})$.

Each rule R1 and R2 is permitted also in dual form, with existential and universal quantifiers interchanged. Rules R0 and R1 say that, for a small space-bounded computation running in n^a time, we may "speed it up" by *guessing* a list $C_1, \ldots, C_{n^x}$ of configurations of the machine at n^x points in time, then *for all* configurations C_i on the list, verifying that the machine simulated from C_i reaches the configuration C_{i+1}, in n^{a-x} steps. The rule R2 is a simple consequence of a translation (a.k.a. "padding") lemma: if NTIME$(n) \subseteq$ DTS(n^c), then for all $a \geq 1$, coNTIME(n^a) and NTIME(n^a) are in DTS(n^{ca}). More background can be found in the excellent surveys by Van Melkebeek [15], [16].

Definition: A *refutation* $\mathcal{D}$ consists of a sequence of lines of the form

$${}^1(\exists n^{a_1})^{b_2}(\forall n^{a_2})^{b_3} \cdots {}^{b_k}(Qn^{a_k})^{b_{k+1}}\text{DTS}(n^{a_{k+1}})$$

where $a_i, b_i \geq 0$ and "Q" is either "$\forall$" or "$\exists$" depending on whether k is even or odd. The line is said to have k *alternations*. The refutation $\mathcal{D}$ must satisfy:

(a) The first line is ${}^1\text{DTS}(n^a)$.
(b) Each line follows from the proceeding line by one of the above rules.
(c) Only the first and last lines may (possibly) have zero quantifiers.
(d) The last line has the form ${}^1\text{DTS}(n^b)$, with $b < a$.

A $\mathcal{D}$ which satisfies conditions (b) and (c) is called a *derivation*.

B. Simplified rules of inference

As a first step towards simplifying the syntax of refutations and derivations, we define the notion of "h-refutation". An *h-derivation* or *h-refutation* is defined similarly to a derivation or refutation, but with the following changes. First, change the leading superscript "1" in all lines to be a "0". Second, replace rule R0 with rule h-R0 by replacing all three superscripts "1" with "0". In particular, the superscript"$\max\{x,1\}$" is replaced by just "x".[1]

$$\text{h-R0}: \qquad {}^0\text{DTS}(n^a) \ \subseteq\ {}^0(\exists n^x)^x(\forall n^0)^0\text{DTS}(n^{a-x}).$$

Lemma 3: There is an h-refutation if and only if there is a refutation.

The difficult direction of Lemma 3 is the transformation of h-refutations into refutations. The intuition is that by scaling the exponents in an h-refutation by a large multiplicative factor, one can make all exponents greater than 1, and then the h-refutation is easily converted to a refutation by suitably replacing exponents "0" with "1".

Proof: ($\Longleftarrow$) Suppose $\mathcal{D}$ is a refutation. We need to form an h-refutation $\mathcal{D}'$. To form $\mathcal{D}'$, first replace the initial line, ${}^1\text{DTS}(n^a)$, of $\mathcal{D}$ with ${}^0\text{DTS}(n^a)$, and change the initial inference of $\mathcal{D}$ to be an h-R0 inference instead of an R0 inference. To form the rest of $\mathcal{D}'$, follow exactly the same inferences as in $\mathcal{D}$. It is easy to check that this can be done in such a way that each line in $\mathcal{D}'$ has exactly the same form as the corresponding line in $\mathcal{D}$ except that some of the exponents in $\mathcal{D}'$ may be less than the corresponding exponents in $\mathcal{D}$.

($\Longrightarrow$) Let $\mathcal{D}'$ be an h-refutation; we must construct a refutation $\mathcal{D}$. Let $\mathcal{D}'(m)$ denote the result of multiplying all superscripts in $\mathcal{D}'$ by the value $m > 0$. Let the first R2 (slowdown) inference in $\mathcal{D}'$ be the i-th inference in $\mathcal{D}'$. Thus, the first $i-1$ inferences in $\mathcal{D}'$ are speedup inferences, h-R0 or R1. Choose m large enough so that $m > 1/x$ for all values of x used in these first $i-1$ speedup inferences.

In $\mathcal{D}'(m)$, the second through i-th lines have the form

$$ {}^0(\exists n^{a_1})^{b_2}\ldots{}^{b_k}(Qn^0)^0\text{DTS}(n^{a_{k+1}}). \tag{1}$$

This is because rule h-R0 gives a formula of this form, and the speedup rule R1 preserves this form. By choice of m, for all $i \le k$, the values a_i and b_i are > 1 in the lines (1). The next line in $\mathcal{D}'(m)$, inferred by slowdown, has the form

$$ {}^0(\exists n^{a_1})^{b_2}\ldots{}^{b_{k-1}}(Qn^{a_{k-1}})^{b_k}\text{DTS}(n^{\max\{cb_k,ca_{k+1}\}}). \tag{2}$$

Form the refutation $\mathcal{D}$ by modifying $\mathcal{D}'(m)$ as follows. First, in the $i-1$ lines of the form (1), replace "$(Qn^0)^0$" with "$(Qn^0)^1$". Second, on every line, replace the leading superscript "0" with "1".

[1]The "h" stands for "homogeneous", and the key property of an h-derivation is that if all superscripts are multiplied by a fixed positive constant, it remains a valid h-derivation.

It is straightforward to verify that this makes $\mathcal{D}$ a valid refutation. The first $i-1$ inferences are correct since $b_k > 1$ by choice of m. The ith-inference, a slowdown, of the line (2) is also correct, since $b_k > 1$. Finally, the first superscripts b_2 are all ≥ 1: this is true for the first line by choice of m, and the values of b_2 can only increase when they are affected by a speedup R2. Thus the final inference in $\mathcal{D}$ has the form

$$ {}^1(\exists n^{a_1})^{b_2}\text{DTS}(n^{a_2}) \ \subseteq\ {}^1\text{DTS}(n^{\max\{ca_1,cb_2,ca_2\}})$$

with $b_2 \ge 1$ and is a valid instance of R2. ∎

Our second simplification removes all the a_i's, $i = 1,\ldots,k$, from lines in derivations. This is based on two observations: First, $a_i \le b_{i+1}$, for all $i \le k$. This property holds for rule h-R0 and is preserved by R1 and R2. Second, the value of a_k is used only for the slowdown rule R2 in the expression $\max\{cb_k, ca_k, cb_{k+1}, ca_{k+1}\}$. But, as $a_k \le b_{k+1}$, the presence of a_k is superfluous.

This observation lets us simplify the proof system considerably. Our "reduced" system replaces each quantifier $(Qn^{a_i})^{b_{i+1}}$ by just $Q^{b_{i+1}}$. Valid lines in a reduced derivation have the form:

$$ {}^0\exists^{b_1}\forall^{b_2}\exists^{b_3}\ldots{}^{b_{k-1}}Q^{b_{k+1}}\text{DTS}(n^a). \tag{3}$$

for $0 \le b_i$ and $0 \le a$. Now (3) no longer represents a complexity class per se – it is merely a syntactic object. Nonetheless, the reduced system allows us to reason about refutations involving "real" complexity classes. We use "$\vdash$" instead of "$\subseteq$" to indicate derivability in the reduced system. The rules of inference for the reduced system are:

R0′: *(Initialization)*
${}^0\text{DTS}(n^a) \vdash {}^0\exists^0\text{DTS}(n^a)$.

R1′: *(Speedup)*
$\ldots{}^{b_k}\forall^{b_{k+1}}\text{DTS}(n^a) \vdash$
$\quad\ldots{}^{b_k}\forall^{\max\{x,b_{k+1}\}}\exists^{b_{k+1}}\text{DTS}(n^{a-x})$,
where $0 < x \le a$.

R2′: *(Slowdown)*
$\ldots{}^{b_k}\forall^{b_{k+1}}\text{DTS}(n^a) \vdash\ \ldots{}^{b_k}\text{DTS}(n^{\max\{cb_k,cb_{k+1},ca\}})$.

As before, each rule R1′ and R2′ is permitted in dual form, with existential and universal quantifiers interchanged. The rule R0′ has been formulated to have only one quantifier and not incorporate a speedup: this will be convenient later when we discuss c-achievable pairs.

A *reduced refutation* is defined similarly to a refutation, but using $\vdash$ instead of $\subseteq$, with rules R0′–R2′ in place of R0–R2, and must prove ${}^0\text{DTS}(n^a)\vdash{}^0\text{DTS}(n^b)$ for $b < a$.

Lemma 4: There is a reduced refutation (with R0′-R2′) iff there is a refutation (with R0-R2).

Proof: Note that an application of R0′ followed by a use of R1′ can simulate a reduced initial speedup (h-R0) inference:

$$ {}^0\text{DTS}(n^a) \vdash {}^0\exists^x\forall^0\text{DTS}(n^{a-x})$$

The lemma thus follows from Lemma 3 and the above discussion. ■

The rest of the paper will work primarily with reduced derivations and refutations. To simplify terminology, we henceforth use the terms "derivation" and "refutation" to refer to reduced derivations and refutations. The context should always make it clear whether we are referring to the reduced or the original system.

C. Approximate inferences

Definition: Let Ξ and Ξ' be classes represented in the reduced inference system just defined:

$$\Xi = {}^{0}\exists^{b_2}\forall^{b_3}\dots{}^{b_k}Q^{b_{k+1}}\mathrm{DTS}(n^a) \quad (4)$$
$$\Xi' = {}^{0}\exists^{b'_2}\forall^{b'_3}\dots{}^{b'_k}Q^{b'_{k+1}}\mathrm{DTS}(n^{a'}).$$

If Ξ and Ξ' have the same number of alternations, then $\Xi' \le \Xi$ iff $a' \le a$ and $b'_i \le b_i$ for all i.

The class $\Xi + \epsilon$ is defined by the condition $\Xi' = \Xi + \epsilon$ holds iff $a' = a + \epsilon$ and $b'_i = b_i + \epsilon$ for all $i \ge 2$.

Definition: The *weakening* rule of inference allows Ξ to be inferred from Ξ' if $\Xi' \le \Xi$. We use the notation $\Xi \overset{w}{\vdash} \Lambda$ to indicate that there is a derivation of Λ from Ξ in the reduced inference system augmented with the weakening rule. A derivation that includes weakening inferences is called a $\overset{w}{\vdash}$*-derivation*. We reserve the terminology "derivation" and the symbol "$\vdash$" for (reduced) derivations that do not use weakenings.

Lemma 5: Let Ξ, Ξ', Λ, Λ' be classes in the reduced refutation system.

(a) $\Xi \overset{w}{\vdash} \Lambda$ iff there is a $\Lambda' \le \Lambda$ such that $\Xi \vdash \Lambda'$.
(b) If $\Xi \overset{w}{\vdash} \Lambda$ and $\Xi' \le \Xi$, then there is a derivation of $\Xi' \vdash \Lambda'$ for some $\Lambda' \le \Lambda$.

The lemma is readily proved by induction on the number of lines in a derivation with weakening rules. We leave the details to the reader. By part (b) of the lemma we may assume WLOG that derivations (without weakening inferences) never contain lines $\Xi \le \Xi'$ with Ξ preceding Ξ' in the derivation.

We next define a notion of "approximate inference", denoted $\Vdash$. Intuitively, $\Xi \Vdash \Lambda$ means that from Ξ one can derive something as close to Λ as desired.

Definition: We write $\Xi \Vdash \Lambda$ to mean that for all $\epsilon > 0$, there exists a $\delta > 0$ so that $(\Xi + \delta) \overset{w}{\vdash} (\Lambda + \epsilon)$.

Lemma 6: The $\Vdash$ relation is transitive: if $\Xi \Vdash \Lambda$ and $\Lambda \Vdash \Gamma$, then $\Xi \Vdash \Gamma$.

Now let Δ be a "prefix" for a reduced line; that is, $\Delta = {}^{0}\exists^{e_2}\forall^{e_3}\dots{}^{e_\ell}\forall^{e_{\ell+1}}$. (Note there is no "DTS" part to Δ.) For Ξ of the form shown above in (4), we define the concatenation $\Delta\Xi$ to be the reduced line

$${}^{0}\exists^{e_2}\forall^{e_3}\dots{}^{e_\ell}\forall^{e_{\ell+1}}\exists^{b_2}\forall^{b_3}\dots{}^{b_k}\forall^{b_{k+1}}\mathrm{DTS}(n^a).$$

A similar definition of concatenation is used for prefixes Δ with an odd number of quantifiers; in this case, since quantifiers must alternate type, if Ξ begins with an $\exists$ then Δ must begin with a $\forall$, and vice-versa.

Lemma 7: If $\Xi \Vdash \Gamma$, then $\Delta\Xi \Vdash \Delta\Gamma$.

Proof: For $\epsilon > 0$, choose $\delta > 0$ so that there is a $\overset{w}{\vdash}$-derivation $\mathcal{D}$ of $\Gamma + \epsilon$ from $\Xi + \delta$. Without loss of generality, $\delta \le \epsilon$. We claim that that, by prefixing each line in $\mathcal{D}$ with $\Delta + \delta$, we obtain a $\overset{w}{\vdash}$-derivation $\mathcal{D}'$ of $(\Delta + \delta)(\Gamma + \epsilon)$ from $(\Delta + \delta)(\Xi + \delta)$. This is because $\mathcal{D}$ contains no lines with zero quantifiers, and thus the superscript "0" at the beginning of each line has no effect on the validity of $\mathcal{D}$. Since $\delta \le \epsilon$, adding a weakening at the end of $\mathcal{D}'$ makes it a $\overset{w}{\vdash}$-derivation of the line $(\Delta + \epsilon)(\Gamma + \epsilon)$. ■

IV. Achievable derivations

A. Achievability and subsumption

Williams [20] uses proof annotations of **1**'s and **0**'s to indicate sequences of speedups and slowdowns (respectively) in a derivation. We think of **1**'s and **0**'s as being paired up like open and closed parentheses, and define a *balanced* derivation to be a derivation containing only inferences of types R1′ and R2′ for which the corresponding pattern of **1**'s and **0**'s, viewed as parentheses, is properly balanced. Put another way, a derivation is balanced provided the first and last lines have the same number of alternations, and each intermediate line has at least that many alternations. In a balanced derivation, each speedup (a "**1**") is uniquely matched by a later slowdown (a "**0**").

We use the star notation * of regular expressions to construct annotations for derivations. For instance, a derivation of type $(\mathbf{10})^*$ consists of alternating speedup and slowdown inferences. Theorems 11 and 16 will establish what can be achieved with derivations of this type.

Definition: Let $\langle \mu, \nu \rangle$ be a pair such that $\mu \ge 1$ and $0 < \nu$. The pair $\langle \mu, \nu \rangle$ is c*-achievable* provided that, for all values a, b and d satisfying $c\mu b = \nu d$,

$${}^{a}\exists^{b}\mathrm{DTS}(n^d) \Vdash {}^{a}\exists^{\mu b}\mathrm{DTS}(n^{\nu d}). \quad (5)$$

The inference (5) is called a $\langle \mu, \nu \rangle$ *step*. A c-achievable pair $\langle \mu, \nu \rangle$ is called *useful* provided $\nu < 1$.

One subtle, but important, aspect of the definition of c-achievable is that the value of a makes no difference at all. This is because the approximate implication (5) must be based on derivations that satisfy condition (c) of the definition of "derivation" as given at the end of Section III-A. That is, the derivations cannot contain any lines with zero quantifiers, and inspection of the rules R1′ and R2′ shows that the value a cannot influence these derivations.

It is also important to note that c-achievability is defined in terms of $\Vdash$, namely, approximate inference. That is, if

$\langle \mu, \nu \rangle$ is c-achievable, it is only required that the $\langle \mu, \nu \rangle$ step be approximately derivable.

The motivation is that we wish to make ν as small as possible in c-achievable pairs so as to make νd as small as possible. This will needed to find as good a refutation as possible (that is, a refutation for as large a value of c as possible). In particular, the next lemma shows that if $\nu < 1/c$ is c-achievable, then there is a refutation.

Lemma 8: Suppose there is a c-achievable $\langle \mu, \nu \rangle$ with $\nu < 1/c$. Then there exists a refutation.

Proof: We have the following (approximate) refutation:

$$\begin{array}{lcll} {}^0\mathrm{DTS}(n^1) & \vdash & {}^0\exists^0\mathrm{DTS}(n^1) & \text{(initialization)} \\ & \overset{w}{\vdash} & {}^0\exists^{\nu/(c\mu)}\mathrm{DTS}(n^1) & \text{(weakening)} \\ & \Vdash & {}^0\exists^{\nu/c}\mathrm{DTS}(n^{\nu}) & \text{(by a } \langle \mu, \nu \rangle \text{ step)} \\ & \vdash & {}^0\mathrm{DTS}(n^{c\nu}) & \text{(slowdown)} \end{array}$$

With $\nu < 1/c$, we have $c\nu < 1$. By definition of approximate derivation ($\Vdash$), we can hence derive ${}^0\mathrm{DTS}(n^{c\nu+\epsilon})$ from ${}^0\mathrm{DTS}(n^1)$ for all small $\epsilon > 0$. Choosing ϵ so that $c\nu + \epsilon < 1$ gives a refutation. ■

The converse to Lemma 8 will be proved below as Lemma 21; thus there is a refutation if and only if there is an achievable pair $\langle \mu, \nu \rangle$ with $\nu < 1/c$.

Unfortunately, making ν small involves a tradeoff: the $\langle \mu, \nu \rangle$ step (5) increases the value of b to $b' = \mu b$ while decreasing the value of d to $d' = \nu d$. Furthermore, as we shall see, obtaining achievable pairs with smaller values of ν will be done at the cost of requiring larger values of μ.

Definition: An implication

$$\ldots {}^{b_k}Q^{b_{k+1}}\mathrm{DTS}(n^a) \overset{w}{\vdash} \ldots {}^{b_k}Q^{b'_{k+1}}\mathrm{DTS}(n^{a'}) \qquad (6)$$

is *subsumed by* $\langle \mu, \nu \rangle$ provided the implication can be inferred by a weakening, followed by a $\langle \mu, \nu \rangle$ step and then a weakening.

The next two lemmas follow from the definitions.

Lemma 9: The implication (6) is subsumed by $\langle \mu, \nu \rangle$ iff

$$b'_{k+1} \geq \max\{\mu b_{k+1}, \frac{1}{c}\nu a\} \text{ and}$$
$$a' \geq \max\{c\mu b_{k+1}, \nu a\}.$$

Lemma 10: Suppose $\mu \leq \mu'$ and $\nu \leq \nu' < 1$. If $\langle \mu, \nu \rangle$ is c-achievable, then so is $\langle \mu', \nu' \rangle$. If an implication is subsumed by $\langle \mu', \nu' \rangle$, then it is also subsumed by $\langle \mu, \nu \rangle$.

We also need a weaker notion of subsumption, which is defined as follows (compare to Lemma 9).

Definition: The implication (6) is *weakly subsumed* by $\langle \mu, \nu \rangle$ iff $a' \geq \max\{c\mu b_{k+1}, \nu a\}$.

The intuition is that optimal derivations in the proof system are subsumed by c-achievable pairs. However, there are also non-optimal derivations that are only weakly subsumed by a c-achievable pair. As an example, the trivial inference ${}^0\mathrm{DTS}(n^d) \vdash {}^0\mathrm{DTS}(n^d)$ is only weakly subsumed by $\langle 1, 1 \rangle$, or indeed by any c-achievable $\langle \mu, \nu \rangle$.

Recall that $1 < c < 2$. The next lemma, although stated quite differently, is essentially the same as the Conditional Speedup Lemma 6.7 of Williams [19].

Lemma 11: The pair $\langle 1, c-1 \rangle$ is c-achievable, with derivations of type $(\mathbf{10})^*$.

Since $c < 2$, the pair $\langle 1, c-1 \rangle$ is useful.

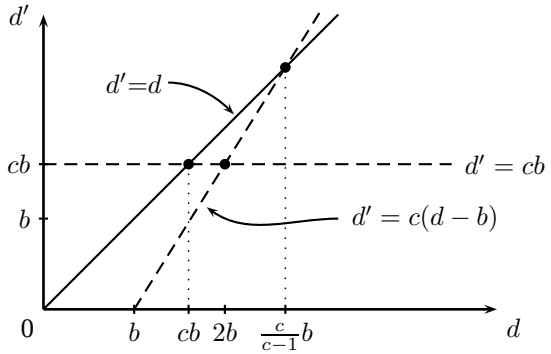
Proof: Let $\Xi = {}^a\exists^b\mathrm{DTS}(n^d)$. If $cb \leq d$, then from Ξ we can derive, by a speedup followed by a slowdown:

$$\begin{array}{lcl} \Xi & \vdash & {}^a\exists^b\forall^b\mathrm{DTS}(n^{d-b}) \\ & \vdash & {}^a\exists^b\mathrm{DTS}(n^{\max\{cb, c(d-b)\}}), \end{array} \qquad (7)$$

where the first step is a speedup with $x = b$. That is, from Ξ we can derive

$${}^a\exists^b\mathrm{DTS}(n^{d'})$$

with $d' = \max\{cb, c(d-b)\}$. The possible values for d' are shown on the following graph.



As shown in the graph, for $d' = \max\{cb, c(d-b)\}$, we have $d' < d$ precisely when $cb < d < \frac{c}{c-1}b$. For $cb \leq d \leq 2b$, we have $d' = cb$. And, for $2b < d < \frac{c}{c-1}b$, we have $d' = c(d-b)$. Thus, depending on the value of d, we have either $d' = cb$ or $\left(\frac{c}{c-1}b - d'\right) = c\left(\frac{c}{c-1}b - d\right)$. Therefore, by repeating the inference pattern $\mathbf{10}$ a finite number of times, we can infer

$${}^a\exists^b\mathrm{DTS}(n^d) \vdash {}^a\exists^b\mathrm{DTS}(n^{cb}), \qquad (8)$$

provided $cb < d < \frac{c}{c-1}b$.

To complete the proof of Lemma 11, we must show that

$${}^a\exists^b\mathrm{DTS}(n^{\frac{c}{c-1}b}) \Vdash {}^a\exists^b\mathrm{DTS}(n^{cb}).$$

Let $\epsilon > 0$, and pick $\delta > 0$ so that $\delta \leq \epsilon/c$ and $\delta < c(2-c)b/(c-1)^2$. By the latter inequality and since $\frac{c}{c-1} > 1$,

$$c(b+\delta) \;<\; \frac{c}{c-1}b + \delta \;<\; \frac{c}{c-1}(b+\delta). \qquad (9)$$

Therefore, we have

$$\begin{array}{rcl} {}^{a+\delta}\exists^{b+\delta}\mathrm{DTS}(n^{\frac{c}{c-1}b+\delta}) & \vdash & {}^{a+\delta}\exists^{b+\delta}\mathrm{DTS}(n^{c(b+\delta)}) \\ & \overset{w}{\vdash} & {}^{a+\epsilon}\exists^{b+\epsilon}\mathrm{DTS}(n^{cb+\epsilon}) \end{array}$$

where the first step follows by a $(\mathbf{10})^*$ derivation as in (8) using (9), and the second step is a weakening, as $c\delta \leq \epsilon$. ■

B. Composition of achievable pairs

We next describe how two c-achievable pairs can be combined (or, "composed") to form another.

Lemma 12: Let $\langle \mu_1, \nu_1 \rangle$ and $\langle \mu_2, \nu_2 \rangle$ be c-achievable. Also suppose $c\nu_1\mu_2 \geq \mu_1$. Set

$$\mu = c\nu_1\mu_2 \tag{10}$$

$$\nu = \frac{c\mu_1\nu_1\nu_2}{\mu_1 + \nu_1\nu_2} \tag{11}$$

Then $\langle \mu, \nu \rangle$ is c-achievable.

The idea of the proof is that a $\langle \mu, \nu \rangle$ step can be achieved by a speedup (R1′) inference, a $\langle \mu_2, \nu_2 \rangle$ step, a slowdown (R2′) inference, and a $\langle \mu_1, \nu_1 \rangle$ step. That is, if B and A are annotations for proofs that approximate a $\langle \mu_1, \nu_1 \rangle$ step and a $\langle \mu_2, \nu_2 \rangle$ step sufficiently well (respectively), then $\mathbf{1}A\mathbf{0}B$ is an annotation for an approximate $\langle \mu, \nu \rangle$ step. However, the final $\langle \mu_1, \nu_1 \rangle$ step should be skipped if $\langle \mu_1, \nu_1 \rangle$ is not useful. The proof can be found in the full version.

The condition $c\nu_1\mu_2 \geq \mu_1$ restricts how c-achievable pairs can be composed. The next lemma shows that the case where this condition fails can be handled by simply letting $\mu = \max\{\mu_1, c\nu_1\mu_2\}$.

Lemma 13: Let $\langle \mu_1, \nu_1 \rangle$ and $\langle \mu_2, \nu_2 \rangle$ be c-achievable. Set

$$\mu = \max\{c\nu_1\mu_2, \mu_1\} \tag{12}$$

$$\nu = \frac{c\mu_1\nu_1\nu_2}{\mu_1 + \nu_1\nu_2}. \tag{13}$$

Then $\langle \mu, \nu \rangle$ is c-achievable.

Proof: If $\mu_1 \leq c\nu_1\mu_2$, then Lemma 12 already implies the result. Otherwise, let $\mu_2' = \mu_1/(c\nu_1)$, so that $\mu_2' > \mu_2$ and $\mu_1 = c\nu_1\mu_2'$. By Lemma 10, $\langle \mu_2', \nu_2 \rangle$ is c-achievable. Thus Lemma 12 applied to the pairs $\langle \mu_1, \nu_1 \rangle$ and $\langle \mu_2', \nu_2 \rangle$ now gives the desired result. ■

To better understand what is happening when we compose $\langle \mu_1, \nu_1 \rangle$ and $\langle \mu_2, \nu_2 \rangle$ to form $\langle \mu, \nu \rangle$, reexpress the two formulas of Lemma 12 as:

$$\frac{1}{\mu} = \frac{1}{c\nu_1}\left(\frac{1}{\mu_2}\right) \tag{14}$$

$$\frac{1}{\nu} = \frac{\nu_1}{\mu_1(c\nu_1 - 1)} - \frac{1}{c\nu_1}\left(\frac{\nu_1}{\mu_1(c\nu_1 - 1)} - \frac{1}{\nu_2}\right) \tag{15}$$

Equations (14) and (15) give an interesting perspective on μ and ν. We may view the pair $\langle \mu_1, \nu_1 \rangle$ as a transformation that acts on the *reciprocal* values $1/\mu_2$ and $1/\nu_2$ to give the values $1/\mu$ and $1/\nu$. Equation (14) shows that the value $1/\mu_2$ is scaled by the factor $1/(c\nu_1)$ to obtain $1/\mu$. In the usual case where $\nu_1 \geq 1/c$, the scale factor is ≤ 1, so $\mu \geq \mu_2$.

Equation (15) shows that the value $1/\nu$ is obtained by contracting $1/\nu_2$ towards a fixed point $\nu_1/(\mu_1(c\nu_1 - 1))$, with the scaling factor for the contraction again equal to $1/(c\nu_1)$. In most cases, $1/\nu_2$ is smaller than this fixed point and we also have the scale factor $1/(c\nu_1) < 1$. In these cases, $1/\nu > 1/\nu_2$, so $\nu < \nu_2$. Getting smaller and smaller values for ν is desirable since, as Lemma 8 showed, our goal is to obtain $\nu < 1/c$ so as to obtain a refutation.

The fixed point for the mapping $\nu_2 \mapsto \nu$ will be denoted by $\tau(\mu_1, \nu_1)$; namely,

$$\tau(\mu_1, \nu_1) = \frac{c\nu_1 - 1}{\nu_1}\mu_1$$

and

$$(\tau(\mu_1, \nu_1))^{-1} = \frac{\nu_1}{(c\nu_1 - 1)\mu_1}.$$

With this notation, we can rewrite equation (15) equivalently as

$$\left((\tau(\mu_1, \nu_1))^{-1} - \frac{1}{\nu}\right) = \frac{1}{c\nu_1}\left((\tau(\mu_1, \nu_1))^{-1} - \frac{1}{\nu_2}\right). \tag{16}$$

This makes it clear how ν is contracting towards $\tau(\mu_1, \nu_1)$.

C. The refutations for $c < 2\cos(\pi/7)$

Suppose that $1 < c < 2\cos(\pi/7)$. Recasting results from Williams [19], we prove there exists a refutation. We begin by recalling a simple characterization of $2\cos(\pi/7)$:

Lemma 14: Let $c \geq 1$. Then

$$\tau(1, c-1) = \frac{c(c-1) - 1}{c - 1} \leq \frac{1}{c}$$

if and only if $c \leq 2\cos(\pi/7)$.

Proof: The inequality holds iff $c^3 - c^2 - 2c + 1 \leq 0$. For $c \geq 1$, this is equivalent to $c \leq 2\cos(\pi/7)$. ■

The following is immediate from Lemma 14 and the fact that $\tau(1, c-1)$ is a fixed point for the composition defined earlier.

Theorem 15 (Williams [19]): There is a refutation for $1 < c < 2\cos(\pi/7)$.

The present work establishes that this is the only situation in which a refutation can exist, under the rules of alternation-trading proofs.

V. The limits of achievable constructions

In this section we argue that, for $1 < c < 2$, no refutation can do better than what is possible using c-achievable pairs, and furthermore that the best c-achievable pairs are $\langle 1, c-1 \rangle$ and the ones that can be obtained by the constructions of Lemmas 12 and 13.

A. Limits on derivations of type $(\mathbf{10})^*$

We start by giving lower bounds on what can be achieved with derivations that follow the $(\mathbf{10})^*$ pattern.

Lemma 16: Any non-empty $(\mathbf{10})^*$ pattern of inferences in a derivation is subsumed by $\langle 1, c-1\rangle$.

Proof: Recall the derivation (7) of type $\mathbf{10}$ that was used in the proof of Lemma 11. We claim that this is the optimal kind of $\mathbf{10}$ inference step. The derivation (7) used a speedup with $x = b$; however, to prove Lemma 16, we must consider a general $\mathbf{10}$ inference with x not necessarily equal to b:

$$\begin{aligned} {}^a\exists^b\mathrm{DTS}(n^d) &\vdash {}^a\exists^{\max\{x,b\}}\forall^b\mathrm{DTS}(n^{d-x}) \\ &\vdash {}^a\exists^{\max\{x,b\}}\mathrm{DTS}(n^{\max\{cx,cb,c(d-x)\}}). \end{aligned}$$

We need to rule out the use of $x \neq b$. First, suppose $x < b$. In this case, we can achieve the same inference by using a weakening to increase the value of d and change the speedup to use $x = b$. Namely,

$$\begin{aligned} {}^a\exists^b\mathrm{DTS}(n^d) &\overset{w}{\vdash} {}^a\exists^b\mathrm{DTS}(n^{d+b-x}) \\ &\vdash {}^a\exists^b\forall^b\mathrm{DTS}(n^{(d+b-x)-b}) \\ &= {}^a\exists^b\forall^b\mathrm{DTS}(n^{d-x}) \\ &\vdash {}^a\exists^b\mathrm{DTS}(n^{\max\{cb,c(d-x)\}}). \end{aligned}$$

Second, suppose $x > b$. In this case, we first use weakening to increase b by $x - b$:

$$\begin{aligned} {}^a\exists^b\mathrm{DTS}(n^d) &\overset{w}{\vdash} {}^a\exists^x\mathrm{DTS}(n^d) \\ &\vdash {}^a\exists^x\forall^x\mathrm{DTS}(n^{d-x}) \\ &\vdash {}^a\exists^x\mathrm{DTS}(n^{\max\{cx,c(d-x)\}}). \end{aligned}$$

Thus any $(\mathbf{10})^*$ pattern of inferences can be replaced by a sequence of operations of the following types: (a) increase d, (b) increase b, and (c) replace d with $\max\{cb, c(d-b)\}$. There is, WLOG, at least one operation of type (c). It is not hard to show that any such sequence of operations is subsumed by $\langle 1, c-1\rangle$. (See the full version for details.) ■

B. Limits on derivations of type $\mathbf{1}A\mathbf{0}B$

The next lemma shows that any balanced derivation that starts with a line of the form $\cdots {}^a\exists^b\mathrm{DTS}(n^d)$ with $d > cb$ does no real work, and can be replaced by a weakening. Thus WLOG, any premiss of a speedup inference has $d > cb$.

Lemma 17: Suppose a balanced derivation starts with the line $\cdots {}^a\exists^b\mathrm{DTS}(n^d)$. Then the last line of the derivation has the form $\cdots {}^a\exists^{b'}\mathrm{DTS}(n^{cb''})$ for some $b'' \geq b' \geq b$.

Thus, if $d \leq cb$, then any non-empty balanced derivation, with first line $\cdots {}^a\exists^b\mathrm{DTS}(n^d)$, is subsumed by $\langle 1, 1\rangle$.

Proof: Throughout the derivation, the superscript after the $\exists$ stays equal to b or becomes larger. (This is because speedup steps can not decrease the superscript, and because the derivation is balanced and cannot remove the $\exists$ with a slowdown.) Therefore, the final step in the derivation is a slowdown of the form

$$\cdots {}^a\exists^{b'}\forall^e\mathrm{DTS}(n^f) \vdash \cdots {}^a\exists^{b'}\mathrm{DTS}(n^{\max\{cb',ce,cf\}}).$$

Letting $b'' = \max\{b, e, f\}$, this proves the lemma. ■

In keeping with the intuition that $\langle \mu_1, \nu_1\rangle$ is a transformation acting on $\langle \mu_2, \nu_2\rangle$, we sometimes express the conditions (10) and (11), or the equivalent (14) and (15), with a mapping notation:

$$\langle \mu_1, \nu_1\rangle : \langle \mu_2, \nu_2\rangle \mapsto \langle \mu, \nu\rangle.$$

This notation is used only when $\mu_1 \leq c\nu_1\mu_2$. Otherwise, we will occasionally express that (12) and (13) hold by writing

$$\langle \mu_1, \nu_1\rangle : \langle \mu_2, \nu_2\rangle \mapsto^{\max} \langle \mu, \nu\rangle.$$

Note that the "$\mapsto^{\max}$" notation makes no restriction on whether μ_1 is larger than $c\nu_1\mu_2$.

The next lemma is our main technical tool putting limitations on how derivations are formed from c-achievable pairs. Informally, it states that any balanced derivation with a **1**/**0** annotation of the form $\mathbf{1}A\mathbf{0}B$ with A and B balanced can be subsumed by the composition of the subderivation A and the subderivation B, where "composition" is in the sense of composition of pairs $\langle \mu_i, \nu_i\rangle$ as used in Lemmas 12 and 13.

Lemma 18: Let a balanced derivation $\mathcal{D}$ have the annotation $\mathbf{1}A\mathbf{0}B$, where A and B are balanced **1**/**0**-patterns. Suppose that the subderivation corresponding to A is weakly subsumed by $\langle \mu_2, \nu_2\rangle$. Further suppose that the subderivation corresponding to B is non-empty and subsumed (respectively, weakly subsumed) by $\langle \mu_1, \nu_1\rangle$. Then the entire derivation $\mathcal{D}$ is subsumed (respectively, weakly subsumed) by a pair $\langle \mu, \nu\rangle$ such that either

$$\langle \mu_1, \nu_1\rangle : \langle \mu_2, \nu_2\rangle \mapsto^{\max} \langle \mu, \nu\rangle, \tag{17}$$

or

$$\langle 1, 1\rangle : \langle \mu_2, \nu_2\rangle \mapsto \langle \mu, \nu\rangle. \tag{18}$$

On the other hand, if B is empty, then the derivation $\mathcal{D}$ is weakly subsumed by the $\langle \mu, \nu\rangle$ given by (18).

The lemma is stated for derivations $\mathcal{D}$ that contain only speedup and slowdown inferences, and no weakenings. However, by the proof of Lemma 5 and the definition of subsumption, it also holds for derivations that contain weakenings. In this case, the weakenings in the derivation do not contribute to the pattern of **0**'s and **1**'s for the derivation.

Proof: The derivation starts with $\Xi = \cdots {}^a\exists^b\mathrm{DTS}(n^d)$, and ends with a line $\Delta = \cdots {}^a\exists^{x'}\mathrm{DTS}(n^{u'})$ (or, dually, with $\forall$ in place of $\exists$). The prefix "$\cdots$" never changes during the balanced derivation, so we henceforth suppress it in the notation. The first inference of the $\mathbf{1}A\mathbf{0}B$ derivation is a speedup,

$${}^a\exists^b\mathrm{DTS}(n^d) \vdash {}^a\exists^{\max\{x,b\}}\forall^b\mathrm{DTS}(n^{d-x}).$$

We claim that WLOG we have $x \geq b$. This is proved just as in the proof of Lemma 16. Namely, if $x < b$, just add a weakening inference to the beginning to derive

$$ {}^a\exists^b\mathrm{DTS}(n^d) \;\models^w\; {}^a\exists^b\mathrm{DTS}(n^{d+b-x}) \vdash {}^a\exists^b\forall^b\mathrm{DTS}(n^{d-x}). $$

This means there is a $\mathbf{1}A\mathbf{0}B$ derivation $\mathcal{D}'$ of Δ from ${}^a\exists^b\mathrm{DTS}(n^{d+b-x})$. Thus, it suffices to prove the lemma assuming that the first speedup inference uses $x \geq b$; this will prove that $\langle \mu, \nu \rangle$ subsumes $\mathcal{D}'$ and hence subsumes $\mathcal{D}$.

The $\mathbf{1}A\mathbf{0}$ portion of the derivation $\mathcal{D}$ consists of a speedup, then a subderivation with the annotation A that is weakly subsumed by $\langle \mu_2, \nu_2 \rangle$, and then a slowdown:

$$\begin{array}{rcll} {}^a\exists^b\mathrm{DTS}(n^d) & \vdash & {}^a\exists^x\forall^b\mathrm{DTS}(n^{d-x}) & \text{(by speedup)} \\ & \vdots & \vdots & \text{(weakly subsumed by } \langle \mu_2, \nu_2 \rangle) \\ & \vdash & {}^a\exists^x\forall^y\mathrm{DTS}(n^z) & \\ & \vdash & {}^a\exists^x\mathrm{DTS}(n^u) & \text{(by slowdown)} \quad (19) \end{array}$$

where $u = \max\{cx, cy, cz\}$ and where, by the weak subsumption by $\langle \mu_2, \nu_2 \rangle$,

$$ z \geq \max\{c\mu_2 b, \nu_2(d-x)\}. $$

Suppose B is empty in the derivation, so ${}^a\exists^x\mathrm{DTS}(n^u)$ is the last line of the $\mathbf{1}A\mathbf{0}B$ derivation. By $u \geq cz$ and $u \geq cx$, we have $u \geq c(c\mu_2)b$ and $u \geq \max\{cx, c\nu_2(d-x)\}$. The value $\max\{cx, c\nu_2(d-x)\}$ is minimized with $x = \nu_2 d/(1+\nu_2)$ and therefore $u \geq c\nu_2 d/(1+\nu_2)$. Thus, if B is empty, the derivation $\mathcal{D}$ is weakly subsumed by the pair $\langle \mu, \nu \rangle$ with $\mu = c\mu_2$ and $\nu = \frac{c\nu_2}{1+\nu_2}$. This is the same as defining μ and ν by $\langle 1, 1 \rangle : \langle \mu_2, \nu_2 \rangle \mapsto \langle \mu, \nu \rangle$.

Now assume B is non-empty. We claim that we may assume WLOG $c\mu_2 b \leq \nu_2(d-x)$. If this does not hold, we argue similarly to how we showed that $x \geq b$ WLOG, and prove that we can increase the value of d to $x + \frac{c\mu_2}{\nu_2}b$. Namely, let $d' = x + \frac{c\mu_2}{\nu_2}b > d$, and replace the $\mathbf{1}A\mathbf{0}$ portion of $\mathcal{D}$ with the following inferences:

$$\begin{array}{rcll} {}^a\exists^b\mathrm{DTS}(n^d) & \models^w & {}^a\exists^b\mathrm{DTS}(n^{d'}) & \text{(weakening)} \\ & \vdash & {}^a\exists^x\forall^b\mathrm{DTS}(n^{d'-x}) & \text{(speedup)} \\ & = & {}^a\exists^x\forall^b\mathrm{DTS}(n^{c\mu_2 b/\nu_2}) & \\ & \Vdash & {}^a\exists^x\forall^{\mu_2 b}\mathrm{DTS}(n^{c\mu_2 b}) & \text{(by a } \langle \mu_2, \nu_2 \rangle \text{ step)} \\ & = & {}^a\exists^x\forall^y\mathrm{DTS}(n^z) & \text{(where } y = \mu_2 b \text{ and } z = c\mu_2 b) \\ & \vdash & {}^a\exists^x\mathrm{DTS}(n^u) & \text{(slowdown)} \end{array}$$

In this case, we still have $z \geq \max\{c\mu_2 b, \nu_2(d-x)\}$. Modifying $\mathcal{D}$ in this way leaves the first and last lines of the derivation intact, so if we prove this modified derivation is subsumed by a pair $\langle \mu, \nu \rangle$ it certainly follows that $\mathcal{D}$ is also subsumed by the same pair.

It thus follows that we can assume WLOG that

$$ b \leq x \leq d - \frac{c\mu_2}{\nu_2}b \qquad (20) $$

with the derivation $\mathcal{D}$ having the annotation $\mathbf{1}A\mathbf{0}B$, possibly with A representing a $\langle \mu_2, \nu_2 \rangle$ step and a weakening.

In the line (19) at the end of the $\mathbf{1}A\mathbf{0}$ part of the derivation, we must have $u \geq cz \geq c\nu_2(d-x)$. Picking up from line (19), the "B" part of the derivation derives

$$ {}^a\exists^x\mathrm{DTS}(n^u) \vdash {}^a\exists^{x'}\mathrm{DTS}(n^{u'}). $$

Since this part is weakly subsumed by $\langle \mu_1, \nu_1 \rangle$, we have

$$ u' \geq \max\{c\mu_1 x, c\nu_1\nu_2(d-x)\}. \qquad (21) $$

If B is also (non-weakly) subsumed by $\langle \mu_1, \nu_1 \rangle$, then

$$ x' \geq \max\{\mu_1 x, \nu_1\nu_2(d-x)\}. \qquad (22) $$

We claim that we can assume without loss of generality that either (i) $x = b$ and $\mu_1 x > \nu_1\nu_2(d-x)$ or (ii) $x \geq b$ and $\mu_1 x \leq \nu_1\nu_2(d-x)$. To prove this, suppose $\mu_1 x > \nu_1\nu_2(d-x)$ and $x > b$. (Recall that we already have $x \geq b$.) Then, we can modify the $\mathbf{1}A\mathbf{0}B$ derivation by decreasing the value of x to get a stronger derivation. The value of x can be decreased until either $x = b$ or $\mu_1 x = \nu_1\nu_2(d-x)$ so that either (i) or (ii) holds.

If case (i) applies, we have $x = b$ and $\mu_1 b \geq \nu_1\nu_2(d-b)$. This gives

$$ (\mu_1 + \nu_1\nu_2)b \geq \nu_1\nu_2 d. \qquad (23) $$

Multiplying (20) by $\nu_1\nu_2$ gives

$$ \nu_1\nu_2 d \geq (\nu_1\nu_2 + c\mu_2\nu_1)b. \qquad (24) $$

The last two equations imply $\mu_1 \geq c\nu_1\mu_2$. The bound (21) with $x \geq b$ implies that $u' \geq c\mu_1 b$. This, plus (23), implies $u' \geq \frac{c\mu_1\nu_1\nu_2}{\mu_1+\nu_1\nu_2}d$. Thus the entire derivation $\mathcal{D}$ is weakly subsumed by $\langle \mu, \nu \rangle$ with

$$\begin{array}{rcl} \mu & = & \mu_1 \;=\; \max\{\mu_1, c\nu_1\mu_2\} \\ \nu & = & \dfrac{c\mu_1\nu_1\nu_2}{\mu_1 + \nu_1\nu_2} \end{array}$$

If B is (non-weakly) subsumed by $\langle \mu_2, \nu_2 \rangle$, then similar reasoning using (22) in place of (21) gives a lower bound on x' and proves that the derivation $\mathcal{D}$ is also (non-weakly) subsumed by $\langle \mu, \nu \rangle$.

If case (i) does not apply, then (ii) $\mu_1 x \leq \nu_1\nu_2(d-x)$ and $x \geq b$. In particular, $(\mu_1 + \nu_1\nu_2)x \leq \nu_1\nu_2 d$, so

$$ x \leq \frac{\nu_1\nu_2}{\mu_1 + \nu_1\nu_2}d \quad \text{and} $$

$$ d - x \geq \frac{\mu_1}{\mu_1 + \nu_1\nu_2}d. \qquad (25) $$

From (20), we get $d - x \geq \frac{c\mu_2}{\nu_2}b$, whence

$$ \nu_1\nu_2(d-x) \geq c\nu_1\mu_2 b. \qquad (26) $$

By (ii), we get $\nu_1\nu_2(d-x) \geq \mu_1 b$. This fact and inequalities (21), (25) and (26) imply that

$$ u' \geq \max\left\{c\mu_1 b, (c^2\nu_1\mu_2)b, \frac{c\mu_1\nu_1\nu_2}{\mu_1 + \nu_1\nu_2}d\right\}. $$

Therefore, the entire derivation $\mathcal{D}$ is weakly subsumed by the pair $\langle \mu, \nu \rangle$, where $\mu = \max\{\mu_1, c\nu_1\mu_2\}$ and $\nu = \frac{c\mu_1\nu_1\nu_2}{\mu_1+\nu_1\nu_2}$. If B was (non-weakly) subsumed by $\langle \mu_2, \nu_2 \rangle$, then, by similar reasoning using (22), $\mathcal{D}$ is also (non-weakly) subsumed by $\langle \mu, \nu \rangle$. This completes the proof of Lemma 18. ■

C. Characterization of achievable pairs

In this section we prove that every balanced derivation is subsumed by some c-achievable pair, and we give a small list of operations that suffice to form all c-achievable pairs.

The earlier constructions used the following five methods for constructing c-achievable pairs:

(A) $\langle 1, c-1 \rangle$ is c-achievable.

(B) Suppose $\langle \mu_1, \nu_1 \rangle$ and $\langle \mu_2, \nu_2 \rangle$ are c-achievable and $\mu_1 \leq c\nu_1\mu_2$. Then $\langle \mu, \nu \rangle$ is c-achievable, where

$$\langle \mu_1, \nu_1 \rangle : \langle \mu_2, \nu_2 \rangle \mapsto \langle \mu, \nu \rangle.$$

(C) Suppose $\langle \mu_1, \nu_1 \rangle$ and $\langle \mu_2, \nu_2 \rangle$ are c-achievable and $\mu_1 > c\nu_1\mu_2$. Then $\langle \mu, \nu \rangle$ is c-achievable, where

$$\langle \mu_1, \nu_1 \rangle : \langle \mu_2, \nu_2 \rangle \mapsto^{\max} \langle \mu, \nu \rangle.$$

(D) If $\langle \mu_2, \nu_2 \rangle$ is c-achievable, then so is $\langle \mu, \nu \rangle$, where

$$\langle 1, 1 \rangle : \langle \mu_2, \nu_2 \rangle \mapsto \langle \mu, \nu \rangle.$$

(E) If $\langle \mu, \nu \rangle$ is c-achievable and $\mu' \geq \mu$ and $1 \geq \nu' \geq \nu$, then $\langle \mu', \nu' \rangle$ is c-achievable.

(Constructions (B) and (C) are defined separately since we will later show that the constructions (C) are not needed.) A pair $\langle \mu, \nu \rangle$ is called an ABCD-*pair* if it can be shown to be c-achievable by the operations (A)-(D).

Theorem 19: Any balanced non-empty derivation $\mathcal{D}$ starting with a line with at least one alternation, is weakly subsumed by some ABCD-pair.

One more simple lemma is needed to prove Theorem 19:

Lemma 20: Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be balanced derivations with the first line of $\mathcal{D}_2$ the same as the last line of $\mathcal{D}_1$. If $\mathcal{D}_1$ is subsumed by the c-achievable pair $\langle \mu, \nu \rangle$, then the concatenation $\mathcal{D}_1\mathcal{D}_2$ is also subsumed by $\langle \mu, \nu \rangle$.

Proof: Let $\mathcal{D}_1$ begin with the line $\cdots {}^a\exists^b\text{DTS}(n^d)$, and end with the line $\cdots {}^a\exists^{b'}\text{DTS}(n^{d'})$. By the subsumption assumption, letting $f = \max\{\mu b, \frac{1}{c}\nu d\}$, we have $b' \geq f$ and $d' \geq cf$. Now, by Lemma 17, the last line of $\mathcal{D}_2$ is of the form ${}^a\exists^{b''}\text{DTS}(n^{d''})$, with $b'' \geq b' \geq f$ and $d'' \geq cb' \geq cf$. That is, $\mathcal{D}_1\mathcal{D}_2$ is also subsumed by $\langle \mu, \nu \rangle$. ■

The proof of Theorem 19 is by induction on the complexity of the derivation $\mathcal{D}$. Since $\mathcal{D}$ is balanced, its first inference is a speedup, and there is later a matching slowdown. That is, $\mathcal{D}$ has the annotation $\mathbf{1}A\mathbf{0}B$ where A and B are balanced patterns of $\mathbf{0}$'s and $\mathbf{1}$'s. If A is empty, then the first two lines of $\mathcal{D}$ are inferred by a $\mathbf{10}$ pattern and hence by Lemma 16 is subsumed by $\langle 1, c-1 \rangle$. Therefore, by Lemma 20, all of $\mathcal{D}$ is also subsumed by $\langle 1, c-1 \rangle$. Now suppose A is non-empty. The induction hypothesis is that the subderivations of $\mathcal{D}$ corresponding to A and B are both weakly subsumed by ABCD-pairs. It follows immediately from Lemma 18 that $\mathcal{D}$ is also weakly subsumed by some ABCD-pair. This concludes the proof of Theorem 19.

D. Characterizing refutations

We can now characterize for which values of $c > 1$ refutations exist, in terms of what pairs are c-achievable.

Lemma 21: Fix $c \geq 1$. There is a refutation if and only if there is some ABCD-pair $\langle \mu, \nu \rangle$ with $\nu < 1/c$. Furthermore, there is a refutation if and only if there is a c-achievable pair with $\nu < 1/c$.

Proof: By Theorem 19, any refutation has the form

$$\begin{array}{lcll} {}^0\text{DTS}(n^1) & \vdash & {}^0\exists^0\text{DTS}(n^1) & \text{Initialization} \\ & \vdots & \vdots & \text{(weakly subsumed by } \langle \mu, \nu \rangle\text{)} \\ & \vdash & {}^0\exists^a\text{DTS}(n^d) & \\ & \vdash & {}^0\text{DTS}(n^{\max\{ca,cd\}}) & \text{Slowdown} \end{array}$$

with $\max\{ca, cd\} < 1$, for an ABCD-pair $\langle \mu, \nu \rangle$. The definition of weak subsumption implies $d \geq \nu$, thus $\nu < 1/c$.

Conversely, every ABCD-pair is c-achievable. And by Lemma 8, if there is c-achievable pair with $\nu < 1/c$, then there is a refutation. ■

VI. Limits on achievable pairs

The previous section reduced the question of whether there exists a refutation to the question of whether there is a c-achievable pair $\langle \mu, \nu \rangle$ with $\nu < 1/c$. It was further shown that only ABCD-pairs need be considered. We shall show, in fact, that only ABE-pairs need to be considered; namely, that any c-achievable pair is subsumed by some ABE-pair.

Definition: The ABE-*pairs* (respectively, AB-*pairs*) are the pairs that can be obtained by operations (A), (B) and (E) (respectively, by (A) and (B)).

A pair $\langle \mu, \nu \rangle$ is *subsumed* by $\langle \mu', \nu' \rangle$ if $\mu' \leq \mu$ and $\nu' \leq \nu$.

Lemma 22: Every ABCD-pair is an ABE-pair.

Proof: The proof of Lemma 13 shows that any use of rule (C) can be replaced by rule (E) followed by rule (B). Since $\langle 1, c-1 \rangle$ subsumes $\langle 1, 1 \rangle$, rule (D) is unnecessary. ■

Corollary 23: Fix $c \geq 1$. There is a refutation if and only if there is some ABE-pair $\langle \mu, \nu \rangle$ with $\nu < 1/c$.

Recall from Section IV-B the definition of τ:

$$\tau(\mu, \nu) \;=\; \frac{c\nu - 1}{\nu}\mu \;=\; \left(c - \frac{1}{\nu}\right)\mu.$$

As we showed, the action of $\langle \mu_1, \nu_1 \rangle$ on $\langle \mu_2, \nu_2 \rangle$ produces $\langle \mu, \nu \rangle$ with ν obtained by "reciprocally contracting" ν_2 towards $\tau(\mu_1, \nu_1)$. The next lemma shows that either $\tau(\mu_1, \nu_1)$ is sufficient for obtaining a refutation or it only causes τ values to increase.

Lemma 24: Suppose $\tau(\mu_1,\nu_1) \geq 1/c$ and $\langle \mu_1,\nu_1\rangle : \langle \mu_2,\nu_2\rangle \mapsto \langle \mu,\nu\rangle$. Then $\tau(\mu,\nu) \geq \tau(\mu_2,\nu_2)$.

Proof: Note that $\frac{1}{\nu} = \frac{1}{c\nu_1\nu_2} + \frac{1}{c\mu_1}$. We have

$$\begin{aligned}
\tau(\mu,\nu) &= \left(c - \frac{1}{\nu}\right)\mu \;=\; \left(c - \frac{1}{c\nu_1\nu_2} - \frac{1}{c\mu_1}\right) c\nu_1\mu_2 \\
&= c^2\nu_1\mu_2 - \frac{\mu_2}{\nu_2} - \frac{\mu_2\nu_1}{\mu_1} \\
&= \left(c\mu_2 - \frac{\mu_2}{\nu_2}\right) + \left(c^2\nu_1\mu_2 - c\mu_2 - \frac{\mu_2\nu_1}{\mu_1}\right) \\
&= \tau(\mu_2,\nu_2) + \left(c\frac{(c\nu_1-1)\mu_1}{\nu_1} - 1\right)\frac{\nu_1\mu_2}{\mu_1} \\
&= \tau(\mu_2,\nu_2) + (c\tau(\mu_1,\nu_1) - 1)\,\frac{\nu_1\mu_2}{\mu_1} \\
&\geq \tau(\mu_2,\nu_2),
\end{aligned}$$

where the last inequality follows from $\tau(\mu_1,\nu_1) \geq 1/c$. ■

Theorem 25: Fix $c \geq 1$. There is a refutation if and only if $c < 2\cos(\pi/7)$.

Proof: Theorem 15 shows that if $c < 2\cos(\pi/7)$, then there is a refutation. For the converse, suppose $c \geq 2\cos(\pi/7)$. We claim that any ABE-pair $\langle \mu,\nu\rangle$ has

$$\tau(\mu,\nu) \geq \tau(1,c-1) \geq 1/c \quad \text{and} \quad \nu > \tau(1,c-1) \geq 1/c. \tag{27}$$

The claim is proved by induction on the number of steps used to derive the ABE-pair. The base case for the induction is $\langle \mu,\nu\rangle = \langle 1, c-1\rangle$. Then, since $c \geq 2\cos(\pi/7)$, we have $\nu = c-1 > 1/c$. Also, $\tau(1,c-1) \geq 1/c$ by Lemma 14. The induction step splits into two cases depending on whether $\langle \mu,\nu\rangle$ is derived by an (E)-operation or a (B)-operation. If it is derived by an (E)-operation (subsumption), then the inequalities of (27) follow immediately from the induction hypothesis and monotonicity. If $\langle \mu,\nu\rangle$ is derived by a (B)-operation, the first inequality of (27) follows from Lemma 24. For the second inequality, observe that by equation (16), if $c\nu_1 > 1$ and $\langle \mu_1,\nu_1\rangle : \langle \mu_2,\nu_2\rangle \mapsto \langle \mu,\nu\rangle$, then ν is between ν_2 and $\tau(\mu_1,\nu_1)$. This proves the claim.

It follows by Corollary 23 that if $c \geq 2\cos(\pi/2)$, there is no proof of a refutation. ■

Theorem 1 is an immediate corollary of Lemma 4 and Theorem 25.

REFERENCES

[1] S. Aaronson and A. Wigderson, "Algebrization: A new barrier in complexity theory," *ACM Transactions on Computation Theory*, vol. 1, no. 1, 2009.

[2] T. Baker, J. Gill, and R. Solovay, "Relativizations of the P=?NP question," *SIAM Journal on Computing*, vol. 4, pp. 431–442, 1975.

[3] S. Buss and R. Williams, "Limits on alternation-trading proofs for time-space lower bounds," ECCC, Tech. Rep. TR11-031, March 2011, submitted for publication.

[4] S. Diehl and D. van Melkebeek, "Time-space lower bounds for the polynomial-time hierarchy on randomized machines," *SIAM Journal on Computing*, vol. 36, pp. 563–594, 2006.

[5] L. Fortnow, "Nondeterministic polynomial time versus nondeterministic logarithmic space: Time-space tradeoffs for satisfiability," in *Proc. IEEE Conference on Computational Complexity (CCC)*, 1997, pp. 52–60.

[6] L. Fortnow, R. Lipton, D. van Melkebeek, and A. Viglas, "Time-space lower bounds for satisfiability," *J. Association for Computing Machinery*, vol. 52, no. 6, pp. 835–865, 2005.

[7] L. Fortnow and D. van Melkebeek, "Time-space tradeoffs for nondeterministic computation," in *Proc. IEEE Conference on Computational Complexity (CCC)*, 2000, pp. 2–13.

[8] R. Impagliazzo, V. Kabanets, and A. Kolokolova, "An axiomatic approach to algebrization," in *Proc. 41st ACM Symp. on Theory of Computing (STOC'09)*, 2009, pp. 695–704.

[9] R. Kannan, "Towards separating nondeterminism from determinism," *Mathematical Systems Theory*, vol. 17, pp. 29–45, 1984.

[10] R. Lipton and A. Viglas, "On the complexity of SAT," in *Proc. 40th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 1999, pp. 459–464.

[11] A. Nash, R. Impagliazzo, and J. B. Remmel, "Universal languages and the power of diagonalization," in *IEEE Conference on Computational Complexity (CCC'03)*, 2003, pp. 337–346.

[12] V. A. Nepomnjaščiĭ, "Rudimentary predicates and Turing computations," *Dokl. Akad. Nauk SSSR*, vol. 195, pp. 282–284, 1970, english translation in *Soviet Math. Dokl.* 11 (1970) 1462–1465.

[13] A. A. Razborov and S. Rudich, "Natural proofs," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 24–35, 1997.

[14] I. Tourlakis, "Time-space tradeoffs for SAT and related problems," *Journal of Computer and System Sciences*, vol. 63, no. 2, pp. 268–287, 2001.

[15] D. van Melkebeek, "Time-space lower bounds for NP-complete problems," in *Current Trends in Theoretical Computer Science*. World Scientific, 2004, pp. 265–291.

[16] ——, "A survey of lower bounds for satisfiability and related problems," *Foundations and Trends in Theoretical Computer Science*, vol. 2, no. 3, pp. 197–303, 2007.

[17] D. van Melkebeek and R. Raz, "A time lower bound for satisfiability," *Theoretical Computer Science*, vol. 348, pp. 311–320, 2005.

[18] R. Williams, "Inductive time-space lower bounds for SAT and related problems," *Computational Complexity*, vol. 15, no. 4, pp. 433–470, 2006.

[19] ——, "Time-space tradeoffs for counting NP solutions modulo integers," *Computational Complexity*, vol. 17, no. 2, pp. 179–219, 2008.

[20] ——, "Alternation-trading proofs, linear programming, and lower bounds," 2009, submitted for publication. A shorter extended abstract appeared in *Proc. 27th Intl. Symp. on Theory of Computings (STACS 2010)*, DOI: 10.4230/LIPIcs.STACS.2010.2494, available from http://stacs-conf.org.