Electronic Colloquium on Computational Complexity, Report No. 31 (2011)

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

Samuel R. Buss*
Department of Mathematics
University of California, San Diego
La Jolla, CA 92093-0112, USA
sbuss@math.ucsd.edu

Ryan Williams†
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120
rrwilliams@gmail.com

March 8, 2011

Abstract

This paper characterizes alternation trading based proofs that satisfiability is not in the time and space bounded class DTISP(n^c, n^ϵ), for various values $c < 2$ and $\epsilon < 1$. We characterize exactly what can be proved in the $\epsilon = 0$ case with currently known methods, and prove the conjecture of Williams that $c = 2\cos(\pi/7)$ is optimal for this. For time-space tradeoffs and lower bounds on satisfiability, we give a theoretical and computational analysis of the alternation trading proofs for $0 < \epsilon < 1$.

1 Introduction

This paper addresses lower bounds for simulating non-determinism with time- and space-bounded deterministic algorithms. We concentrate on lower bounds for deterministic algorithms that solve the satisfiability problem SAT; however, the results also imply analogous time-space lower bounds for many other NP-complete problems (see [11]).

Let DTISP(n^c, n^ϵ) 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 [5, 2, 6, 4, 3, 8, 1, 12, 13, 14], have established better and better non-trivial constant lower bounds on the

*Supported in part by NSF grant DMS-0700533.
†Supported by the Josef Raviv Memorial Fellowship.

ISSN 1433-8092

values c and ϵ for which SAT $\in$ DTISP(n^c, n^ϵ). Surveys of these and other results are given by van Melkebeek [9, 10] but, loosely speaking, these lower bounds have all been obtained by combining a "speedup" technique of Nepomnjascii [7] with an assumption such as SAT $\in$ DTISP(n^c, n^ϵ) in order to obtain a contradiction. Williams [13, 14] gave a formal definition of these proof methods, which he calls "alternation trading proofs", and gave improved time-space lower bounds for deterministic algorithms for SAT. He also designed computer programs that search for optimal alternation trading proofs, and based on these results obtained further alternation trading proofs. He conjectured in [14] that the proofs found by the computer searches are optimal for alternation trading proofs.

The present paper examines more carefully the possible alternation trading proofs for establishing time and space lower bounds on algorithms for SAT. Our first main result is that, for the case of $\epsilon = 0$, the lower bounds obtained by Williams [13, 14] are in fact optimal, at least within the present framework of alternation trading proofs. This proves Williams' conjecture. As part of this, we give some surprising simplifications of alternation trading proofs by characterizing the possible alternation trading proofs with "achievable pairs".

Our second main result is to establish detailed simultaneous time and space lower bounds on deterministic algorithms using alternation trading proofs. Prior work on time-space tradeoffs includes [8, 1, 3, 13, 14]. In particular, [3] showed that if SAT $\in$ DTISP(n^c, n^ϵ) then $n + \epsilon \geq 1.573$, and [13, 14] improved this to $n + \epsilon \geq 2\cos(\pi/7)$ as well as gave better bounds for specific numeric values of c and ϵ. The present paper substantially generalizes results obtained by Williams [13, 14], by establishing that arbitrary alternation trading proofs can be characterized in terms of "achievable triples". We give extensive computer-based searches for achievable triples, aided by theorems on how the search space can be pruned. As a consequence, we are able to find better numerical values for the time-space tradeoffs than those found by [14]. Our computer-based proofs always succeed in establishing either the existence or non-existence of alternation trading proofs. Therefore, our numerical values for simultaneous time and space bounds are the best that can be obtained with presently-known methods for alternation trading proofs.

The time and time-space lower bounds in the present paper are all stated for the problem SAT. As remarked above, they also apply to many other NP-complete problems. In addition, by [13], our lower bounds also apply to the classes 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.

Let $\mathrm{DTS}(n^c)$ denote the class $\mathrm{DTISP}(n^c, n^0)$, that is the set of languages accepted by a deterministic Turing machine with runtime $n^{c+o(1)}$ using space $n^{o(1)}$. Williams [13, 14] proved that SAT $\notin \mathrm{DTS}(n^c)$ for $c < 2\cos(\pi/2)$. For this, he used bounded quantifier notations 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, the notation $(\exists n^2)^1\mathrm{DTS}(n^3)$ denotes the class of languages that are 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.

For his alternation trading proofs, Williams introduced a general framework for establishing lower bounds based on a formal proof system of inference rules that act on bounded quantifier notations for classes. The first kind of inference rules are "speedup" rules that use Nepomnjascii's method of decreasing runtime at the cost of adding alternation(s). The second kind of inference rules, called "slowdown" inferences, use the assumption that $\mathrm{NTIME}(n) \subseteq \mathrm{DTS}(n^c)$ (which follows from SAT $\in \mathrm{DTS}(n^c)$) 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 $\mathrm{DTS}(n^a) \subseteq \mathrm{DTS}(n^{a'})$ for constants $a > a' > 0$. Williams showed SAT $\notin \mathrm{DTS}(n^c)$ for any $c < 2\cos(\pi/7) \approx 1.8019$ by using alternation trading proofs. Based partly on his computer-based searches, he further conjectured that the constant $2\cos(\pi/7)$ is the best that can be obtained with the formalized speedup and slowdown rules.

We prove this conjecture as Theorem 1. The inference rules R0-R2 are defined below in Section 2.

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

The proof of Theorem 1 is based on a new detailed 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. We give methods for generating c-achievable pairs, and prove that these pairs exactly characterize the refutations can be approximated by alternation trading proofs.

Williams used binary strings, called "proof annotations", to 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 "**1010**" represents the sequence of inferences speedup-slowdown-speedup-slowdown. 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 proved these patterns of inferences, as i increases, give contradictions for c arbitrarily close to $2\cos(\pi/7)$, and he conjectured they are the best possible inference patterns. We prove this below as part of proving Theorem 1.

The second half of the paper considers lower bounds on $\mathrm{DTISP}(n^c, n^\epsilon)$ 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 when there are alternation trading proofs that satisfiability is not in $\mathrm{DTISP}(n^c, n^\epsilon)$. Instead, we have to use 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 triple together "dual-subsume" a third triple, as well as a related notion of "multisubsumption". In the end, this allows the computer-based searches to search for quite long proofs. In addition, the computer-based search has always been successful either in finding that an alternation trading refutation exists, or in completely exhausting the search space and thus showing that there is no such refutation.

The outline of the paper is as follows. Section 2 introduces the speedup and slowdown rules, and the notion of alternation trading proofs of $\mathrm{SAT} \notin \mathrm{DTS}(n^c)$. We then introduce some considerably simplified notions of alternation trading proofs, called "h-derivations" and "reduced" derivations, along with some simplified versions of the speedup and slowdown rules, R0′–R2′. Section 3 introduces the notion of approximate inferences, and the notion of a "c-achievable pair", by which is meant that that certain kinds of results can be approximately proved (achieved) with alternation proofs. Section 4 puts limits on what kinds of pairs are c-achievable. Section 5 proves a certain kind of normal form on c-achievable pairs, and completes the proof of Theorem 1. Section 6 turns to time-space tradeoffs and introduces the different systems of alteration trading inferences for $\mathrm{DTISP}(n^c, n^\epsilon)$, including the "reduced" inference system. It also introduces the notion of approximate inferences for DTISP derivations. Section 7 defines achievable triples, and gives methods for generating achievable triples. Section 8 gives the theoretical results needed for our computer-based search for achievable triples, and reports the numerical results of the searches.

Section 9 establishes that our rules for generating (c, ϵ)-achievable triples exactly characterize the possible DTISP refutations.

We review notation and results from earlier work as needed; however, we presume a certain level of familiarity with prior work such as can be found in Williams [14].

2 Rules of inference for DTS

2.1 Basic rules of inference for DTS bounds

Fix, henceforth, a value $c > 1$. The goal is to prove a contradiction from the assumption $\mathrm{SAT} \in \mathrm{DTS}(n^c)$, thereby of course proving that $\mathrm{SAT} \notin \mathrm{DTS}(n^c)$. The contradiction is proved by an *alternation trading proof* using the following rules R0–R2. As shown in [14], it suffices to give an alternation trading proof of $\mathrm{DTS}(n^a) \subseteq \mathrm{DTS}(n^b)$ for some $b < a$. The alternation trading proof is a sequence of containments, starting with the set ${}^1\mathrm{DTS}(n^a)$ 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 [14]. The ellipses "$\cdots$" indicate an arbitrary (possibly empty) quantifier prefix.

R0**:** *Initial speedup:*

$$ {}^1\mathrm{DTS}(n^a) \subseteq {}^1(\exists n^x)^{\max\{x,1\}}(\forall n^0){}^1\mathrm{DTS}(n^{a-x}), $$

where $0 < x \leq a$.

R1**:** *Speedup:*

$$ \cdots {}^{b_k}(\forall n^{a_k})^{b_{k+1}}\mathrm{DTS}(n^{a_{k+1}}) \subseteq \cdots {}^{b_k}(\forall n^{\max\{x,a_k\}})^{\max\{x,b_{k+1}\}}(\exists n^0)^{b_{k+1}}\mathrm{DTS}(n^{a_{k+1}-x}), $$

where $0 < x \leq a_{k+1}$.

R2**:** *Slowdown:*

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

Each rule R1 and R2 is permitted also in dual form, with existential and universal quantifiers interchanged.

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

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

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 $^1\mathrm{DTS}(n^a)$.

(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 $^1\mathrm{DTS}(n^b)$, with $b < a$.

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

2.2 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".

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

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.

Lemma 2 *Fix $c > 1$. There is an h-refutation if and only if there is a refutation.*

The difficult direction of Lemma 2 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, ${}^1\mathrm{DTS}(n^a)$, of $\mathcal{D}$ with ${}^0\mathrm{DTS}(n^a)$, 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

$$ {}^0(\exists n^{a_1})^{b_2}(\forall n^{a_2})^{b_3}\cdots{}^{b_k}(Qn^0)^0\mathrm{DTS}(n^{a_{k+1}}). \tag{1}$$

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 $i \leq k$, 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

$$ {}^0(\exists n^{a_1})^{b_2}(\forall n^{a_2})^{b_3}\cdots{}^{b_{k-1}}(Qn^{a_{k-1}})^{b_k}\mathrm{DTS}(n^{\max\{cb_k,ca_{k+1}\}}). \tag{2}$$

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".

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

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

with $b_2 \geq 1$ and is a valid instance of R2. □

For our second simplification of the syntax of derivations, we shall remove all the a_i's, $i = 1,\ldots,k$, from lines in derivations. This is based on two observations: First, $a_i \leq b_{i+1}$, for all $i \leq k$. 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, being $\leq b_{k+1}$, the presence of a_k is superfluous.

This allows us to simplify the format of lines and rules of inference considerably with a "reduced" inference system. The reduced system replaces each quantifier $(Qn^{a_i})^{b_{i+1}}$ by just $Q^{b_{i+1}}$. The valid lines in a reduced derivation have the form:

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

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

R0′: *Initialization:*

$$^0\mathrm{DTS}(n^a) \vdash {}^0\exists^0\mathrm{DTS}(n^a).$$

R1′: *Speedup:*

$$\ldots{}^{b_k}\forall^{b_{k+1}}\mathrm{DTS}(n^a) \vdash \ldots{}^{b_k}\forall^{\max\{x,b_{k+1}\}}\exists^{b_{k+1}}\mathrm{DTS}(n^{a-x}),$$

where $0 < x \leq a$.

R2′: *Slowdown:*

$$\ldots{}^{b_k}\forall^{b_{k+1}}\mathrm{DTS}(n^a) \vdash \ldots{}^{b_k}\mathrm{DTS}(n^{\max\{cb_k,cb_{k+1},ca\}}).$$

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 $^0\mathrm{DTS}(n^a)\vdash{}^0\mathrm{DTS}(n^b)$ for $b < a$.

Lemma 3 *Fix $c > 1$. 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:

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

The lemma thus follows from Lemma 2 and the above discussion. □

The rest of the paper will work primarily with reduced derivations and refutations. In order 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.

2.3 Approximate inferences

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

$$\Xi \;=\; {}^0\exists^{b_2}\forall^{b_3}\ldots{}^{b_k}Q^{b_{k+1}}\mathrm{DTS}(n^a) \tag{4}$$

and

$$\Xi' \;=\; {}^0\exists^{b'_2}\forall^{b'_3}\ldots{}^{b'_k}Q^{b'_{k+1}}\mathrm{DTS}(n^{a'}).$$

Suppose, as indicated, that Ξ and Ξ' have the same number of alternations. Then $\Xi' \leq \Xi$ iff $a' \leq a$ and $b'_i \leq b_i$ 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 $i \geq 2$.

Definition The *weakening* rule of inference allows Ξ to be inferred from Ξ' if $\Xi' \leq \Xi$. We use the notation $\Xi \models^w \Lambda$ to indicate that there is a derivation of Λ from Ξ in the reduced inference system augmented with the weakening rule.

A derivation that is allowed to containing weakening inferences will be called a $\models^w$*-derivation*. We reserve the terminology "derivation" and the symbol "$\vdash$" for (reduced) derivations that do not use weakenings.

Lemma 4 *Let Ξ, Ξ', Λ, Λ' be classes in the reduced refutation system.*

(a) $\Xi \models^w \Lambda$ *iff there is a* $\Lambda' \leq \Lambda$ *such that* $\Xi \vdash \Lambda'$.

(b) *If* $\Xi \models^w \Lambda$ *and* $\Xi' \leq \Xi$, *then there is a derivation of* $\Xi' \vdash \Lambda'$ *for some* $\Lambda' \leq \Lambda$.

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 without loss of generality that derivations (without weakening inferences) never contain lines $\Xi \leq \Xi'$ 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 $(\Xi + \delta) \stackrel{w}{\vdash} (\Lambda + \epsilon)$.

Lemma 5 *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:

$$\Delta \;=\; {}^{0}\exists^{e_2}\forall^{e_3} \ldots {}^{e_\ell}\forall^{e_{\ell+1}}.$$

(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

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

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 6 *If* $\Xi \Vdash \Gamma$*, then* $\Delta\Xi \Vdash \Delta\Gamma$*.*

Proof For $\epsilon > 0$, choose $\delta > 0$ so that there is a $\stackrel{w}{\vdash}$-derivation $\mathcal{D}$ of $\Gamma + \epsilon$ from $\Xi + \delta$. Without loss of generality, $\delta \leq \epsilon$. We claim that that, by prefixing each line in $\mathcal{D}$ with $\Delta + \delta$, we obtain a $\stackrel{w}{\vdash}$-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 $\delta \leq \epsilon$, adding a weakening at the end of $\mathcal{D}'$ makes it a $\stackrel{w}{\vdash}$-derivation of the line $(\Delta + \epsilon)(\Gamma + \epsilon)$. $\Box$

3 Achievable derivations

3.1 Achievability and subsumption

Williams [14] 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. In other words, 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 inference (a "**1**") is uniquely matched by a later slowdown derivation (a "**0**").

We use the star notation * of regular expressions to construct annotations for derivations. For instance, a derivation of type **(10)*** consists of alternating speedup and slowdown inferences. Theorems 10 and 15 will establish what can be achieved with derivations of this type.

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

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

The inference displayed is called a $\langle \mu, \nu \rangle$ *step*. The 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 2.1. That is to say, 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 to say, 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 derivations so as to make νd as small as possible. This will needed to find as good a refutation as possible (that is to say, 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 7 *Fix $c > 1$. Suppose there is a c-achievable pair $\langle \mu, \nu \rangle$ with $\nu < 1/c$. Then there exists a refutation.*

Proof We have the following (approximate) refutation:

$$\begin{array}{lll} {}^0\mathrm{DTS}(n^1) & \vdash & {}^0\exists^0\mathrm{DTS}(n^1) \qquad \text{Initialization} \\ & \overset{w}{\models} & {}^0\exists^{\nu/(c\mu)}\mathrm{DTS}(n^1) \qquad \text{Weakening} \\ & \Vdash & {}^0\exists^{\nu/c}\mathrm{DTS}(n^\nu) \qquad \text{By a } \langle \mu, \nu \rangle \text{ step} \\ & \vdash & {}^0\mathrm{DTS}(n^{c\nu}) \qquad \text{Slowdown} \end{array}$$

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

The converse to Lemma 7 will be proved below as Lemma 20; 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

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

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 are immediate from the definitions. To prove Lemma 8, note that a $\langle \mu, \nu \rangle$ step was defined with the requirement that $c\mu b = \nu d$: if this equality does not hold, a weakening can be used to increase one of b or d so that $\langle \mu, \nu \rangle$ step can be applied.

Lemma 8 *The implication* (6) *is subsumed by $\langle \mu, \nu \rangle$ iff*

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

Lemma 9 *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 8).

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 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 $\langle \mu, \nu \rangle$.

3.2 Derivations of type (10)*

We continue to fix a value of c with $1 < c < 2$. The next lemma, although stated quite differently, is essentially the same as the Conditional Speedup Lemma 6.7 of Williams [13].

Lemma 10 *The pair $\langle 1, c-1 \rangle$ is c-achievable, with derivations of type* **(10)***.

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

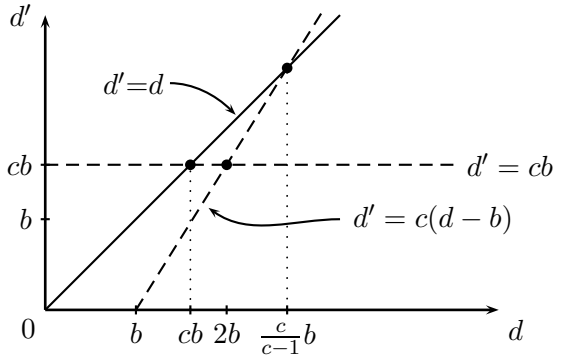
Proof Let $\Xi = {}^a\exists^b\mathrm{DTS}(n^d)$. If $cb \leq d$, then from Ξ we can derive, by a speedup followed by a slowdown:

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

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 **10** a finite number of times, we can infer

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

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

To complete the proof of Lemma 10, we must show that

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

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

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

Therefore, we have

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

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

3.3 Composition of achievable pairs

We next describe how two c-achievable pairs can be combined (or, "composed") to form another c-achievable pair. The next lemma is in some sense equivalent to the construction behind Lemma 6.8 of [13], but is stated in a quite different and more general form.

Lemma 11 *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 for 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.

Proof Let $d = c\mu\frac{1}{\nu}b$ and $\Xi = {}^a\exists^b\mathrm{DTS}(n^d)$. Note that $c\mu b = \nu d$; we have to show that $\Xi \Vdash {}^a\exists^{\mu b}\mathrm{DTS}(n^{c\mu b})$. Let $x = \frac{1}{\mu_1}\mu b = \frac{1}{\mu_1}c\nu_1\mu_2 b$. Since $c\nu_1\mu_2 \geq \mu_1$, we have $x \geq b$. Therefore, by a speedup inference,

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

We have

$$d \;=\; \frac{c\mu}{\nu}b \;=\; c(c\nu_1\mu_2)\left(\frac{\mu_1+\nu_1\nu_2}{c\mu_1\nu_1\nu_2}\right)b \;=\; c\left(\frac{\mu_2}{\nu_2}+\frac{\mu_2\nu_1}{\mu_1}\right)b,$$

whence

$$d - x \;=\; c\frac{\mu_2}{\nu_2}b \;>\; 0.$$

Thus, by the c-achievability of $\langle \mu_2, \nu_2 \rangle$,

$$\Xi \;\Vdash\; {}^a\exists^x\forall^{\mu_2 b}\mathrm{DTS}(n^{\nu_2(d-x)}) \;=\; {}^a\exists^x\forall^{\mu_2 b}\mathrm{DTS}(n^{c\mu_2 b}). \tag{12}$$

The construction now splits into two cases depending on whether $x \leq c\mu_2 b$. First, consider the case $x \leq c\mu_2 b$. Note that this case always applies if $\langle \mu_1, \nu_1 \rangle$ is useful since then $\mu_1 \geq 1$ and $\nu_1 < 1$. Since $x \leq c\mu_2 b$, a slowdown inference applied to (12) gives

$$\Xi \;\Vdash\; {}^a\exists^x\mathrm{DTS}(n^{c\nu_2(d-x)}).$$

A simple calculation shows $c\mu_1 x = \nu_1(c\nu_2(d-x))$. Hence, by the c-achievability of $\langle \mu_1, \nu_1 \rangle$ and the transitivity of $\Vdash$,

$$\Xi \;\Vdash\; {}^a\exists^{\mu_1 x}\mathrm{DTS}(n^{c\mu_1 x}) \;=\; {}^a\exists^{\mu b}\mathrm{DTS}(n^{c\mu b}) \;=\; {}^a\exists^{\mu b}\mathrm{DTS}(n^{\nu d}).$$

On the other hand, suppose $x \geq c\mu_2 b$. Picking up from (12), with a slowdown and a weakening, we obtain,

$$\Xi \;\Vdash\; {}^a\exists^x\mathrm{DTS}(n^{cx}) \;\overset{w}{\vDash}\; {}^a\exists^{\mu_1 x}\mathrm{DTS}(n^{c\mu_1 x}) \;=\; {}^a\exists^{\mu b}\mathrm{DTS}(n^{\nu d}).$$

This proves Lemma 11. □

The condition $c\nu_1\mu_2 \geq \mu_1$ puts a restriction on how c-achievable pairs can be combined by Lemma 11. The next lemma shows that the case where this condition fails can be handled by the simple expedient of letting $\mu = \max\{\mu_1, c\nu_1\mu_2\}$.

Lemma 12 *Let $\langle \mu_1, \nu_1 \rangle$ and $\langle \mu_2, \nu_2 \rangle$ be c-achievable. Set*

$$\mu = \max\{c\nu_1\mu_2, \mu_1\} \tag{13}$$

$$\nu = \frac{c\mu_1\nu_1\nu_2}{\mu_1 + \nu_1\nu_2}. \tag{14}$$

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

Proof If $\mu_1 \leq c\nu_1\mu_2$, then Lemma 11 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 9, $\langle \mu_2', \nu_2 \rangle$ is c-achievable. Thus Lemma 11 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 formulas (10) and (11) as follows:

$$\frac{1}{\mu} = \frac{1}{c\nu_1}\left(\frac{1}{\mu_2}\right) \tag{15}$$

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

Equations (15) and (16) give an interesting perspective on how μ and ν are defined. They allow us to view the pair $\langle \mu_1, \nu_1 \rangle$ as being 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 (15) 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 (16) 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 7 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,

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

With this notation, we can rewrite equation (16) as

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

or equivalently as

$$\left((\tau(\mu_1,\nu_1))^{-1} - \frac{1}{\nu}\right) \;=\; \frac{1}{c\nu_1}\left((\tau(\mu_1,\nu_1))^{-1} - \frac{1}{\nu_2}\right). \tag{17}$$

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

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 (15) and (16), 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 $\mu_1 \le c\nu_1\mu_2$. Otherwise, we will occasionally express that (13) and (14) 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 $c\nu_1\mu_2$.

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

Suppose that $1 < c < 2\cos(\pi/7)$. Recasting results from Williams [13], we prove there exists a refutation. We begin by recalling a simple characterization of $2\cos(\pi/7)$:

Lemma 13 *Let $c \ge 1$. Then*

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

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

Proof The inequality holds iff $c^3 - c^2 - 2c + 1 \le 0$. For $c \ge 1$, this is equivalent to $c \le 2\cos(\pi/7)$, see Williams [13]. □

Theorem 14 (Williams [13]) *There is a refutation for $1 < c < 2\cos(\pi/7)$.*

Proof Define $\mu_0 = 1$ and $\nu_0 = c - 1$. Define $\langle \mu_{k+1}, \nu_{k+1} \rangle$ inductively by $\langle \mu_0, \nu_0 \rangle : \langle \mu_k, \nu_k \rangle \mapsto \langle \mu_{k+1}, \nu_{k+1} \rangle$, so that $\langle \mu_{k+1}, \nu_{k+1} \rangle$ is the composition of $\langle 1, c-1 \rangle$ and $\langle \mu_k, \nu_k \rangle$. Namely,

$$\mu_{k+1} \;=\; c\nu_0\mu_k \qquad \text{and} \qquad \nu_{k+1} \;=\; \frac{c\mu_0\nu_0\nu_k}{\mu_0 + \nu_0\nu_k}.$$

By Lemma 10, $\langle \mu_0, \nu_0 \rangle$ is c-achievable. Thus, if $\nu_0 < 1/c$, then by Lemma 7, there is a refutation. An easy calculation shows that for $c > 1$, we have $\nu_0 = c - 1 < 1/c$ provided $c < (1 + \sqrt{5})/2 \approx 1.618 < 2\cos(\pi/7)$.

It remains to consider the case $(1+\sqrt{5})/2 \leq c < 2\cos(\pi/7)$. This implies that $c\nu_0 \geq 1$. Arguing inductively on k, since ν_0 is $\geq 1/c$, the values of μ_k are increasing with $\mu_k = c\nu_0\mu_{k-1} \geq \mu_{k-1} \geq \mu_0$, which by Lemma 11 implies that $\langle \mu_k, \nu_k \rangle$ is c-achievable.

By (16) and since $c\nu_0 > 1$, the values of ν_k are contracting towards the limit value

$$\tau(\mu_0, \nu_0) \;=\; \frac{\mu_0(c\nu_0 - 1)}{\nu_0} \;=\; \frac{c(c-1) - 1}{c - 1}$$

By Lemma 13, this value is $< 1/c$. Thus, for sufficiently large k, we have $\nu_k < 1/c$. Q.E.D. Theorem 14

It is easy to check that the annotations for the alternation trading proofs described above are the patterns X_i described in the introduction, and the same as proof annotations introduced by Williams [12].

4 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 11 and 12.

4.1 Limits on derivations of type (10)*

We start by giving lower bounds on what can be achieved with derivations that follow the **(10)*** pattern.

Lemma 15 *Any non-empty* **(10)*** *pattern of inferences in a derivation is subsumed by the c-achievable pair* $\langle 1, c-1 \rangle$.

Proof Recall the derivation (7) of type **10** that was used in the proof of Lemma 10. We claim that this is the optimal kind of **10** inference step. The derivation (7) used a speedup with $x = b$; however, to prove Lemma 15, we must consider a general **10** inference with x not necessarily equal to b:

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

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,

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

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

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

Thus any **(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, w.l.o.g., at least one operation of type (c). Referring to the figure used in the proof of Lemma 10, it is evident that any such sequence of operations is subsumed by $\langle 1, c-1\rangle$. □

4.2 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, without loss of generality, any premiss of a speedup inference has $d > cb$.

Lemma 16 *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. □

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 $\mathbf{1}/\mathbf{0}$ 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 11 and 12.

Lemma 17 *Let a balanced derivation $\mathcal{D}$ have the annotation $\mathbf{1}A\mathbf{0}B$, where A and B are balanced $\mathbf{1}/\mathbf{0}$-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*

$$\langle \mu_1, \nu_1 \rangle : \langle \mu_2, \nu_2 \rangle \mapsto^{\max} \langle \mu, \nu \rangle, \tag{18}$$

or

$$\langle 1, 1 \rangle : \langle \mu_2, \nu_2 \rangle \mapsto \langle \mu, \nu \rangle. \tag{19}$$

On the other hand, if B is empty, then the derivation $\mathcal{D}$ is weakly subsumed by the $\langle \mu, \nu \rangle$ given by (19)*.*

The lemma is stated for derivations $\mathcal{D}$ that contain only speedup and slowdown inferences, and no weakenings. However, by the proof of Lemma 4 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 $\mathbf{0}$'s and $\mathbf{1}$'s for the derivation.

Proof The derivation starts with a line equal to $\Xi = \cdots {}^{a}\exists^{b}\mathrm{DTS}(n^{d})$, 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 w.l.o.g. we have $x \geq b$. This is proved just as in the proof of Lemma 15. Namely, if $x < b$, just add a weakening inference to the beginning to derive

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

In particular, 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 will suffice to prove the lemma under the assumption that the first speedup inference uses $x \geq b$, as 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:

$$ \begin{array}{lll} {}^{a}\exists^{b}\mathrm{DTS}(n^{d}) & \vdash & {}^{a}\exists^{x}\forall^{b}\mathrm{DTS}(n^{d-x}) \qquad \text{- by speedup} \\ & \vdots & \vdots \qquad (\text{weakly subsumed by } \langle \mu_2, \nu_2 \rangle) \\ & \vdash & {}^{a}\exists^{x}\forall^{y}\mathrm{DTS}(n^{z}) \\ & \vdash & {}^{a}\exists^{x}\mathrm{DTS}(n^{u}) \qquad \text{- by slowdown} \end{array} \tag{20} $$

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 the B part of the derivation is empty, 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

$$ \mu = c\mu_2 \qquad \text{and} \qquad \nu = \frac{c\nu_2}{1+\nu_2}. $$

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 w.l.o.g. $c\mu_2 b \leq \nu_2(d-x)$. If this does not hold, we argue similarly to how we showed that $x \geq b$ w.l.o.g., 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:

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

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 with no loss of generality that

$$b \leq x \leq d - \frac{c\mu_2}{\nu_2} b \tag{21}$$

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

In the line (20) 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 (20), the "B" part of the derivation derives

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

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

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

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

$$x' \geq \max\{\mu_1 x, \nu_1\nu_2(d-x)\}. \tag{23}$$

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

$$(\mu_1 + \nu_1\nu_2)b \geq \nu_1\nu_2 d. \tag{24}$$

Multiplying (21) by $\nu_1\nu_2$ gives

$$\nu_1\nu_2 d \geq (\nu_1\nu_2 + c\mu_2\nu_1)b. \tag{25}$$

The last two equations imply $\mu_1 \geq c\nu_1\mu_2$. The bound (22) with $x \geq b$ implies that $u' \geq c\mu_1 b$. This, plus (24), 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

$$\begin{aligned} \mu &= \mu_1 = \max\{\mu_1, c\nu_1\mu_2\} \\ \nu &= \frac{c\mu_1\nu_1\nu_2}{\mu_1 + \nu_1\nu_2} \end{aligned}$$

If B is (non-weakly) subsumed by $\langle \mu_2, \nu_2 \rangle$, then similar reasoning using (23) in place of (22) 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

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

and

$$d - x \geq \frac{\mu_1}{\mu_1 + \nu_1\nu_2}d. \tag{26}$$

From (21), we get $d - x \geq \frac{c\mu_2}{\nu_2}b$, whence

$$\nu_1\nu_2(d - x) \geq c\nu_1\mu_2 b. \tag{27}$$

By (ii), we get $\nu_1\nu_2(d - x) \geq \mu_1 b$. This fact and inequalities (22), (26) and (27) imply that

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

Therefore, the entire derivation $\mathcal{D}$ is weakly subsumed by the pair $\langle \mu, \nu \rangle$

$$\begin{aligned} \mu &= \max\{\mu_1, c\nu_1\mu_2\} \\ \nu &= \frac{c\mu_1\nu_1\nu_2}{\mu_1 + \nu_1\nu_2} \end{aligned}$$

If B was (non-weakly) subsumed by $\langle \mu_2, \nu_2 \rangle$, then, by similar reasoning using (23), $\mathcal{D}$ is also (non-weakly) subsumed by $\langle \mu, \nu \rangle$. Q.E.D. Lemma 17. □

4.3 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 18 *Any balanced non-empty derivation $\mathcal{D}$ starting with a line with at least one alternation, is weakly subsumed by some* ABCD-*pair.*

As we shall show momentarily, Theorem 18 follows easily from Lemmas 15 and 17. First, however, we prove another simple lemma.

Lemma 19 *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 (of Lemma 19) Let $\mathcal{D}_1$ begin with the line $\cdots {}^a\exists^b\mathrm{DTS}(n^d)$, and end with the line $\cdots {}^a\exists^{b'}\mathrm{DTS}(n^{d'})$. 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 16, the last line of $\mathcal{D}_2$ is of the form ${}^a\exists^{b''}\mathrm{DTS}(n^{d''})$, with $b'' \geq b' \geq f$ and $d'' \geq cb' \geq cf$. That is to say, $\mathcal{D}_1\mathcal{D}_2$ is also subsumed by $\langle \mu, \nu \rangle$. □

The proof of Theorem 18 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 **0**'s and **1**'s. If A is empty, then the first two lines of $\mathcal{D}$ are inferred by a **10** pattern and hence by Lemma 15 is subsumed by $\langle 1, c-1 \rangle$. Therefore, by Lemma 19, 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 17 that $\mathcal{D}$ is also weakly subsumed by some ABCD-pair. Q.E.D. Theorem 18 □

4.4 Characterizing refutations

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

Lemma 20 *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 18, any refutation must have the form

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

with $\max\{ca, cd\} < 1$, for some ABCD-pair $\langle \mu, \nu \rangle$. By the definition of weak subsumption, this implies $d \geq \nu$. Thus $\nu < 1/c$.

Conversely, every ABCD-pair is c-achievable. And by Lemma 7, if there is c-achievable pair with $\nu < 1/c$, then there is a refutation. □

5 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 $\langle \mu', \nu' \rangle$ when $\mu' \leq \mu$ and $\nu' \leq \nu$.

Lemma 21 *Every* ABCD-*pair is an* ABE-*pair.*

Proof The proof of Lemma 12 shows that any use of rule (C) can be replaced by a use of rule (E) followed by rule (B). And, since $\langle 1, c-1 \rangle$ subsumes $\langle 1, 1 \rangle$, rule (D) is unnecessary. □

Corollary 22 *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 3.3, 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 23 *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

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

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

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

Theorem 24 *There is a refutation if and only if* $c < 2\cos(\pi/7)$.

Proof Theorem 14 already showed 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

$$\tau(\mu,\nu) \geq \tau(1,c-1) \geq 1/c \quad \text{and} \quad \nu > \tau(1,c-1) \geq 1/c. \tag{28}$$

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 13. 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 (28) follow immediately from the induction hypothesis and monotonicity. On the other hand, if $\langle\mu,\nu\rangle$ is derived by a (B)-operation, the first inequality of (28) follows from Lemma 23. The second inequality follows from the fact that equation (17) showed that 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 has value between ν_2 and $\tau(\mu_1,\nu_1)$. This proves the claim.

It follows by Corollary 22 that if $c \geq 2\cos(\pi/2)$, there is no proof of a refutation. Q.E.D. Theorem 24

Theorem 1 is an immediate corollary of Lemma 3 and Theorem 24.

6 Rules of inference for DTISP

So far we have concentrated on the case of algorithms in DTS(n^a), namely algorithms that use only $n^{o(1)}$ space. The rest of the paper considers the

classes DTISP(n^a, n^e) which are allowed to use $n^{e+o(1)}$ space. In particular, we generalize our earlier results to prove time-space tradeoffs that give lower bounds on the values of c and ϵ for which satisfiability can be computed in DTISP(n^c, n^ϵ). Williams [14] has already proved some tradeoffs for several values of c and ϵ. We will extend these bounds, giving a precise tradeoff, and proving that this tradeoff is optimal for the present-day known rules R0$_\epsilon$-R2$_\epsilon$ given below.

6.1 Basic and reduced rules of inference for DTISP bounds

We henceforth fix values $c > 1$ and $\epsilon \in [0, 1)$ such that $c + \epsilon < 2$. Our goal will be to prove a contradiction from the assumption that SAT $\in$ DTISP(n^c, n^ϵ), based on alternation trading inferences. Generalizing the rules R0-R2 discussed for DTS(n^c), we have the following rules of inference for alternation trading proofs, which were implicitly stated in [14]. (The rule names include a subscript ϵ to distinguish them from the earlier-defined rules, but only R2$_\epsilon$ actually depends on c or ϵ.)

R0$_\epsilon$: *Initial speedup:*

$$^1\text{DTISP}(n^a, n^e) \subseteq {}^1(\exists n^x)^{\max\{x,1\}}(\forall n^0)^{\max\{e,1\}}\text{DTISP}(n^{a-x+e}, n^e),$$

where $e < x \leq a$. (The initial speedup rule will be invoked only with $a = c \cdot e/\epsilon$.)

R1$_\epsilon$: *Speedup:*

$$\begin{aligned}&\ldots {}^{b_k}(\exists n^{a_k})^{b_{k+1}}\text{DTISP}(n^{a_{k+1}}, n^e)\\ &\quad\subseteq \ldots {}^{b_k}(\exists n^{\max\{x,a_k\}})^{\max\{x,b_{k+1}\}}(\forall n^0)^{\max\{b_{k+1},e\}}\text{DTISP}(n^{a_{k+1}-x+e}, n^e),\end{aligned}$$

where $e < x \leq a_{k+1}$.

R2$_\epsilon$: *Slowdown:*

$$\ldots {}^{b_k}(\forall n^{a_k})^{b_{k+1}}\text{DTISP}(n^{a_{k+1}}, n^e) \subseteq \ldots {}^{b_k}\text{DTISP}(n^{ca}, n^{\epsilon a}).$$

where $a = \max\{b_k, a_k, b_{k+1}, a_{k+1}\}$. It is required that $e \leq b_{k+1}$.

As before, each rule R1$_\epsilon$ and R2$_\epsilon$ is permitted in dual form, with existential and universal quantifiers interchanged.

Recall that the class DTISP(n^a, n^e) is defined only for $a \geq e$. The upper bound on x in the speedup rule enforces this condition. The conclusion of

the $\mathrm{R2}_\epsilon$ rule is a class $\mathrm{DTISP}(n^b, n^e)$ where $e = \epsilon \cdot b/c$. For space reasons, we shall on rare occasions use the notation $\mathrm{DTISP}(n^b, n^{\cdots})$ for this class, where it is understood that the omitted space bound is n^e with $e = \epsilon \cdot b/c$.

The concept of a refutation is defined similarly as before. It is now required that the first line have the form ${}^1\mathrm{DTISP}(n^a, n^e)$, and the last line have the form ${}^1\mathrm{DTISP}(n^b, n^f)$ where $b < a$ and $e < f$. Since the last line has $f = (\epsilon/c)b$, it can be required without loss of generality that the first line have $e = (\epsilon/c)a$. For fixed values of c and ϵ, if there is a refutation, then $\mathrm{SAT} \notin \mathrm{DTISP}(n^c, n^\epsilon)$.

We now formulate simplified rules of inference for DTISP. First, the h-$\mathrm{R0}_\epsilon$ rule is defined to be:

$$\text{h-R0}_\epsilon\text{:} \qquad {}^0\mathrm{DTISP}(n^a, n^e) \vdash {}^0(\exists n^x)^x(\forall n^0)^e\mathrm{DTISP}(n^{a-x+e}, n^e),$$

namely, by changing the superscripts 1 in $\mathrm{R0}_\epsilon$ to 0's. A homogeneous refutation, or an h-refutation, is one that uses rules h-$\mathrm{R0}_\epsilon$, $\mathrm{R1}_\epsilon$, and $\mathrm{R2}_\epsilon$. By exactly the same proof as Lemma 2, there is an h-refutation iff there is a refutation.

For the second simplification, we form a *reduced* inference system by getting rid of the superscripts a_j bounding the size of the existentially and universally quantified values. The valid lines in a reduced proof will have the form

$${}^0\exists^{b_1}\forall^{b_2}\exists^{b_3}\cdots{}^{b_k}Q^{b_{k+1}}\mathrm{DTISP}(n^a, n^e).$$

These will be required to satisfy the conditions

$$a \geq e \quad \text{and} \quad b_i \geq e, \text{ for all } i. \tag{29}$$

The rules of inference for the reduced system simplify to:

$\mathrm{R0}'_\epsilon$**:** *Initialization:*

$$ {}^0\mathrm{DTISP}(n^a, n^e) \vdash {}^0\exists^e\mathrm{DTISP}(n^a, n^e).$$

$\mathrm{R1}'_\epsilon$**:** *Speedup:*

$$\begin{aligned} &\cdots{}^{b_k}\exists^{b_{k+1}}\mathrm{DTISP}(n^a, n^e) \\ &\quad\vdash \quad \cdots{}^{b_k}\exists^{\max\{x,b_{k+1}\}}\forall^{\max\{b_{k+1},e\}}\mathrm{DTISP}(n^{a-x+e}, n^e), \end{aligned}$$

where $e < x \leq a$. By (29), this rule can be further simplified by replacing "$\max\{b_{k+1}, e\}$" with just "b_{k+1}".

R2$'_\epsilon$: *Slowdown:*

$$\ldots {}^{b_k}\forall^{b_{k+1}}\mathrm{DTISP}(n^a, n^e) \vdash \ \ldots {}^{b_k}\mathrm{DTISP}(n^{ca'}, n^{\epsilon a'}).$$

where $a' = \max\{b_k, b_{k+1}, a, e\} = \max\{b_k, b_{k+1}, a\}$.

As always, the last two rules are permitted in dual form, with existential and universal quantifiers interchanged.

The analogue of Lemma 3 holds for DTISP proofs, so we have:

Lemma 25 *Fix $c > 1$ and $0 \leq \epsilon < 1$. There is a reduced refutation, with* R0$'_\epsilon$-R2$'_\epsilon$*, iff there is a refutation (that is, with* R0$_\epsilon$-R2$_\epsilon$*).*

Proof The proof of Lemma 25 is similar to that of Lemma 3, but we also need to verify that the conditions of (29) can be required to hold. We in addition need that $e \geq \frac{\epsilon}{c}a$ for all lines in the refutation. These conditions are readily shown by induction on the number of steps in a reduced proof: the only slightly problematic condition is that $b_i \geq e$ in all lines. It is certainly true for the conclusion of a R0$'_\epsilon$ or R1$'_\epsilon$ inference. Consider the conclusion $\ldots {}^{b_k}\mathrm{DTISP}(n^{ca'}, n^{\epsilon a'})$ of a R2$'_\epsilon$ slowdown inference. This is matched by an earlier line $\ldots {}^{b''_k}\mathrm{DTISP}(n^{a''}, n^{e''})$ which is the premise of the matching speedup inference. We have $b_k \geq b''_k$, and have $b''_k \geq e'' \geq \frac{\epsilon}{c}a''$ by the induction hypothesis. Also, without loss of generality, either $ca' < a''$ or $\epsilon a' < e''$, since otherwise the second line is a weakening of the first line and the derivation could be simplified. In either case, it follows that $\epsilon a' < e''$. Thus, since $b_i \geq e''$ for all i, we also have $b_i \geq \epsilon a'$ as desired. □

We henceforth work exclusively with reduced derivations and refutations. It is required that the conditions (29) hold for all lines in reduced derivations and refutations.

As a digression, it is interesting to note that the rules R1$_\epsilon$ and R1$'_\epsilon$ could be relaxed removing the restriction that $e < x \leq a$ replaced with the restriction that $e < x \leq a + e$. The relaxed versions of the rules are:

alt-R1$_\epsilon$: *Speedup (alternate form):*

$$\begin{aligned}&\ldots {}^{b_k}(\exists n^{a_k})^{b_{k+1}}\mathrm{DTISP}(n^{a_{k+1}}, n^e)\\ &\quad\subseteq\ \ \ldots {}^{b_k}(\exists n^{\max\{x', a_k\}})^{\max\{x', b_{k+1}\}}(\forall n^0)^{\max\{b_{k+1}, e'\}}\mathrm{DTISP}(n^{a_{k+1}-x+e}, n^{e'}),\end{aligned}$$

where $e < x \leq a_{k+1} + e$, and where $x' = \min\{x, a_{k+1}\}$ and $e' = \min\{e, a_{k+1} - x + e\}$.

alt-R1$'_\epsilon$: *Speedup (alternate reduced form):*

$$\begin{aligned}&\ldots\, {}^{b_k}\exists^{b_{k+1}}\mathrm{DTISP}(n^a, n^e)\\ &\quad\vdash\quad \ldots\, {}^{b_k}\exists^{\max\{x', b_{k+1}\}}\forall^{\max\{b_{k+1}, e'\}}\mathrm{DTISP}(n^{a-x+e}, n^{e'}),\end{aligned}$$

with the same conditions on x, x', and e' (replacing a_{k+1} with a).

There are two main properties to establish about the alt-R1$_\epsilon$ rule. First, this rule is admissible; that is, it only derives true conclusions. Second, if there is a (reduced) refutation using alt-R1$'_\epsilon$, then there is already a reduced refutation.

To verify that alt-R1$_\epsilon$ is a valid inclusion, consider a predicate P that is in $\mathrm{DTISP}(n^{a_{k+1}}, n^e)$. Suppose $a_{k+1} < x \leq a_{k+1} + e$, so $x' = a_{k+1}$. Let $f = x - a_{k+1} > 0$. The predicate P has runtime $n^{a_{k+1}+o(1)}$. To speedup P, split the computation of P into n^f many "blocks" each with computation time $n^{a_{k+1}+o(1)}/n^f = n^{a_{k+1}-x+e+o(1)}$, and existentially guess the following values for each block boundary: (1) The contents of all memory locations that are needed for the computation P in the immediately preceding or immediately following block, and (2) For each such memory location, the index of the previous block boundary where the same memory location is existentially specified. Guessing these values requires space $n^f \cdot n^{a_{k+1}-f+o(1)} = n^{a_{k+1}+o(1)}$. Then universally choose (1) each block and verify its computation, and (2) each pair of block boundaries and verify the consistency of the values of the memory locations and their indices for the previous boundary where the value was specified. The requires $n^{2f} = n^{o(1)}$ many universal choices, plus deterministic computation time and space of $n^{a_{k+1}-x+e+o(1)}$.

Thus, P is computable with $n^{x'+o(1)}$ existential guesses, followed by $n^{o(1)}$ universal choices, and deterministic computation time $n^{a_{k+1}-x+e+o(1)}$. This establishes the admissibility of the inference rule alt-R1$_\epsilon$. From this, it follows that if there is a refutation using the rules R0$_\epsilon$, alt-R1$_\epsilon$, and alt-R2$_\epsilon$, then $\mathrm{SAT} \notin \mathrm{DTISP}(n^c, n^\epsilon)$. And, by the same reasoning used earlier, there is such a refutation if and only if there is a reduced refutation using alt-R1$'_\epsilon$.

Finally, we claim that alt-R1$'_\epsilon$ inferences can be removed from a reduced refutation. For this, suppose that a refutation contains an alt-R1$'_\epsilon$ inference, which must be later followed by a pair of R2$'_\epsilon$ inferences starting at the same quantifier alternation level:

$$\begin{aligned}&\ldots\, {}^{b_k}\exists^{b_{k+1}}\mathrm{DTISP}(n^{a_{k+1}}, n^e)\\ &\quad\vdash\quad \ldots\, {}^{b_k}\exists^{\max\{x', b_{k+1}\}}\forall^{\max\{b_{k+1}, e'\}}\mathrm{DTISP}(n^{a_{k+1}-x+e}, n^{e'})\\ &\quad\vdots\qquad\qquad\vdots\end{aligned}$$

$$\vdash \quad \ldots {}^{b_k}\exists^{y}\mathrm{DTISP}(n^{b}, n^{f})$$
$$\vdash \quad \ldots {}^{b_k}\mathrm{DTISP}(n^{cb'}, n^{\epsilon b'}).$$

In the first inference, we have $x' = a_{k+1}$ by assumption. In the next to last line, we must have $y \geq \max\{x', b_{k+1}\}$ since exponents on quantifiers can only increase during the derivation. In the final inference, $b' = \max\{y, b, b_k\}$. It follows that $b' \geq \max\{a_{k+1}, b_k, b_{k+1}\}$: therefore, (a strengthened version of) the last line can already be derived from the first line by a single $\mathrm{R2}'_{\epsilon}$-inference. Thus, we have shown that alt-$\mathrm{R1}'_{\epsilon}$-inferences are unnecessary for refutations, and can be eliminated from derivations.

Nonetheless, for technical reasons, we henceforth work only with derivations that do not use the alternate speedup rules.

6.2 Approximate inferences for DTISP proofs

The notion of approximate inferences for DTISP proofs is defined similarly to the definitions given in section 2.3. Suppose Ξ and Ξ' are classes represented in the reduced inference system for DTISP:

$$\Xi \;=\; {}^{0}\exists^{b_2}\forall^{b_3}\ldots {}^{b_k}Q^{b_{k+1}}\mathrm{DTISP}(n^{a}, n^{e}) \tag{30}$$

and

$$\Xi' \;=\; {}^{0}\exists^{b'_2}\forall^{b'_3}\ldots {}^{b'_k}Q^{b'_{k+1}}\mathrm{DTISP}(n^{a'}, n^{e'}).$$

We define $\Xi' \leq \Xi$ iff $a' \leq a$ and $e' \leq e$ and $b'_i \leq b_i$ for all i.

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

The *weakening* rule is defined exactly as before. The notation $\Xi \overset{w}{\vdash} \Lambda$ means there is a derivation of Λ from Ξ in the reduced system augmented with the weakening rule. We henceforth reserve the term "derivation" and the symbol "$\vdash$" for derivations that do not use weakenings.

It is easy to check that, with these definitions, Lemma 4 holds word-for-word for DTISP derivations. In particular, the weakening rule does not make possible any new refutations. Furthermore, we may assume w.l.o.g. that no derivation contains two lines $\Xi \leq \Xi'$ with Ξ preceding Ξ' in the derivation.

The notion of approximate inference, $\Vdash$, is defined exactly as before. Lemma 5 still holds; namely, $\Vdash$ is transitive. Similarly, Lemma 6 also holds for approximate DTISP derivations.

7 Achievable derivations for DTISP

7.1 Achievable triples and subsumption

The notion of a "c-achievable pair" was crucial for understanding the power of refutations for DTS proofs. For inferences involving DTISP, there are space bounds in addition to time bounds; as a consequence, it is necessary to consider *triples* $\langle \mu, \nu, \ell \rangle$ instead of pairs $\langle \mu, \nu \rangle$.

Definition Fix values c and ϵ so that $c > 1 > \epsilon \geq 0$ and $c + \epsilon < 2$. Suppose $\mu \geq 1$ and $0 < \nu < 1$ and $1 \leq \ell \in \mathbb{N}$. Then the triple $\langle \mu, \nu, \ell \rangle$ is (c, ϵ)-achievable provided that, for all values b, d and e satisfying $(c+\epsilon)\mu b = \nu(d + \ell e)$ and $e \leq b \leq d$,

$$ {}^{a}\exists^{b}\mathrm{DTISP}(n^{d}, n^{e}) \;\Vdash\; {}^{a}\exists^{\mu b}\mathrm{DTISP}(n^{c\mu b}, n^{\epsilon\mu b}). \tag{31} $$

The displayed inference is called a $\langle \mu, \nu, \ell \rangle$-*step*. The triple $\langle \mu, \nu, \ell \rangle$ is called *useful* provided that $\nu < (c + \epsilon)/(c + \ell\epsilon)$.

As the nomenclature suggests, we are mostly interested in (c, ϵ)-achievable triples that are useful. In particular, when working with DTISP classes so that $e = (\epsilon/c)d$, usefulness is equivalent to having $c\mu b < d$ and $\epsilon\mu b < e$ in (31).

As before, the value of a makes no difference at all in the definition of achievability, since the derivations that approximate the $\Vdash$-implication of (31) cannot contain any lines with zero alternations. For the same reason, any such derivation must end with a slowdown inference, $\mathrm{R2}'_{\epsilon}$; therefore the right hand side of (31) can be assumed to be of the form $\mathrm{DTISP}(n^{d'}, n^{(\epsilon/c)d'})$ with no loss of generality.

The extra restriction that $b \leq d$ for (31) can be made without any loss of generality. Indeed, the next lemma shows that if $b \geq d$ then the best possible next inference is a slowdown. In particular, there is no need to apply a (c, ϵ)-achievable triple when $b \geq d$.

Lemma 26 *Suppose* Ξ *is a line in a refutation of the form* $\cdots {}^{a}\exists^{b}\mathrm{DTISP}(n^{d}, n^{e})$ *with* $b \geq d$. *Then, without loss of generality, the next inference is a slowdown.*

Proof Ξ is followed by some balanced inference pattern (possibly empty) of **1**'s and **0**'s, and then a slowdown that removes the existential quantifier from Ξ. Namely,

$$ \begin{aligned} {}^{a}\exists^{b}\mathrm{DTISP}(n^{d}, n^{e}) \;&\vdash\; {}^{a}\exists^{b'}\mathrm{DTISP}(n^{d'}, n^{e'}) && \text{Balanced inferences} \\ &\vdash\; {}^{a}\mathrm{DTISP}(n^{\max\{ca, cb', cd'\}}, n^{\cdots}) && \text{Slowdown} \end{aligned} $$

where $b' \geq b$. This can be replaced by a single slowdown

$$ {}^{a}\exists^{b}\mathrm{DTISP}(n^{d}, n^{e}) \vdash {}^{a}\mathrm{DTISP}(n^{\max\{ca,cb,cd\}}, n^{\cdots}). $$

Since $b' \geq b \geq d$, this improves the subderivation. □

In seeking alternation trading refutations, our goal is to find a (c,ϵ)-achievable triple with ν and ℓ as small as possible. More precisely, our goal is to have $(c+\ell\epsilon)\nu < (c+\epsilon)/c$. In addition, we shall need to have $c\mu\epsilon < 1$.

Definition We define $\rho(\mu,\nu,\ell) = \dfrac{c(c+\ell\epsilon)\nu}{c+\epsilon}$.

Note that a triple $\langle \mu,\nu,\ell >$ is useful iff it has $\rho(\mu,\nu,\ell) < 1$.

Lemma 27 *Suppose there is a (c,ϵ)-achievable triple $\langle\mu,\nu,\ell\rangle$ with*

$$ c\mu\epsilon \;<\; 1 \qquad \textit{and} \qquad \rho(\mu,\nu,\ell) \;<\; 1. $$

Then there is a refutation.

Proof We have the following (approximate) refutation, where we let $b = \max\{\epsilon, \frac{\nu(c+\ell\epsilon)}{(c+\epsilon)\mu}\}$. Note that $b < 1/c$ since $\mu \geq 1$.

$$
\begin{array}{lcll}
{}^{0}\mathrm{DTISP}(n^{c}, n^{\epsilon}) & \vdash & {}^{0}\exists^{\epsilon}\mathrm{DTISP}(n^{c}, n^{\epsilon}) & \text{Initialization} \\
& \models^{w} & {}^{0}\exists^{b}\mathrm{DTISP}(n^{c}, n^{\epsilon}) & \\
& \Vdash & {}^{0}\exists^{\mu b}\mathrm{DTISP}(n^{c\mu b}, n^{\epsilon\mu b}) & \\
& & & \text{By the } (c,\epsilon)\text{-achievable } \langle\mu,\nu,\ell\rangle \\
& \vdash & {}^{0}\mathrm{DTISP}(n^{c^2\mu b}, n^{\epsilon c\mu b}) & \text{Slowdown}
\end{array}
$$

Since $c^2\mu b < c$, this suffices to prove the lemma. □

The notions of "subsume" and "weakly subsume" carry over to DTISP derivations in the expected way.

Definition An implication

$$ \cdots {}^{b_k}Q^{b_{k+1}}\mathrm{DTISP}(n^{a}, n^{e}) \models^{w} \cdots {}^{b_k}Q^{b'_{k+1}}\mathrm{DTISP}(n^{a'}, n^{e'}) \qquad (32) $$

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

The next two lemmas follow immediately from the definitions.

Lemma 28 *The implication* (32) *is subsumed by* $\langle \mu, \nu, \ell \rangle$ *iff*

$$b'_{k+1} \geq \max\left\{\mu b_{k+1}, \frac{\nu(a+\ell e)}{c+\epsilon}\right\}$$

and

$$a' \geq \max\left\{c\mu b_{k+1}, \frac{c\nu(a+\ell e)}{c+\epsilon}\right\}, \tag{33}$$

and

$$e' \geq \max\left\{\epsilon\mu b_{k+1}, \frac{\epsilon\nu(a+\ell e)}{c+\epsilon}\right\}.$$

Lemma 29 *Suppose* $\mu \leq \mu'$, $\nu \leq \nu'$, *and* $\ell \leq \ell'$. *If* $\langle \mu, \nu, \ell \rangle$ *is* (c, ϵ)*-achievable, then so is* $\langle \mu', \nu', \ell' \rangle$. *Any implication subsumed by* $\langle \mu', \nu', \ell' \rangle$ *is also subsumed by* $\langle \mu, \nu, \ell \rangle$.

Definition The implication (32) is *weakly subsumed* by $\langle \mu, \nu, \ell \rangle$ iff the second inequality of (33) holds, that is, iff a' satisfies the lower bound of (33).

7.2 Derivations of type (10)*

The next lemma is a generalization of Lemma 10.

Lemma 30 *The triple* $\langle 1, c+\epsilon-1, 1 \rangle$ *is* (c, ϵ)*-achievable with* **(10)*** *derivations.*

This triple is useful since $c + \epsilon < 2$.

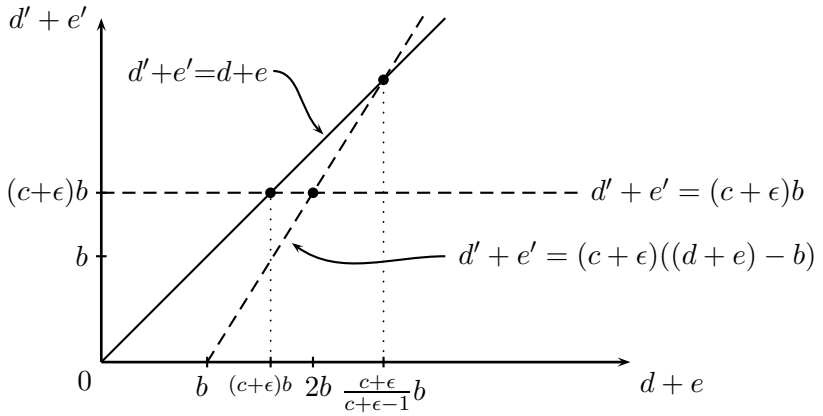
Proof Let $\Xi = {}^{a}\exists^{b}\mathrm{DTISP}(n^d, n^e)$ with $e \leq b \leq d$. A speedup with parameter $x = b$, followed by a slowdown gives

$$\Xi \vdash {}^{a}\exists^{b}\forall^{b}\mathrm{DTISP}(n^{d-b+e}, n^e) \vdash {}^{a}\exists^{b}\mathrm{DTISP}(n^{\max\{cb, c(d-b+e)\}}, n^{\cdots}).$$

Thus, from Ξ we can derive ${}^{a}\exists^{b}\mathrm{DTISP}(n^{d'}, n^{e'})$ with $d' = \max\{cb, c(d-b+e)\}$ and $e' = \max\{\epsilon b, \epsilon(d-b+e)\}$. In particular,

$$d' + e' \;=\; \max\{(c+\epsilon)b, (c+\epsilon)((d+e)-b)\}.$$

The graph below shows the value of $d' + e'$ as a function of $d + e$.



The situation is identical to that of the proof of Lemma 10, except that we are now analyzing how the value $d+e$ changes (instead of d), and the value $c+\epsilon$ replaces the value c. Therefore, if $d+e=\frac{c+\epsilon}{c+\epsilon-1}b$,

$$ {}^a\exists^b\mathrm{DTISP}(n^d,n^e) \;\Vdash\; {}^a\exists^b\mathrm{DTISP}(n^{d'},n^{e'}) $$

where $d'+e'=(c+\epsilon)b$. Since also $e'=(\epsilon/c)d'$, we have $d'=cb$ and $e'=\epsilon b$.

That is to say, if $(c+\epsilon)b=(c+\epsilon-1)(d+e)$ then

$$ {}^a\exists^b\mathrm{DTISP}(n^d,n^e) \;\Vdash\; {}^a\exists^b\mathrm{DTISP}(n^{cb},n^{\epsilon b}) $$

This shows that $\langle 1, c+\epsilon-1, 1\rangle$ is (c,ϵ)-achievable. □

7.3 Composition of achievable triples

The next lemma describes how two (c,ϵ)-achievable triples can be composed to form a third (c,ϵ)-achievable triple.

Lemma 31 *Let $\langle\mu_1,\nu_1,\ell_1\rangle$ and $\langle\mu_2,\nu_2,\ell_2\rangle$ be (c,ϵ)-achievable triples. Let*

$$ \mu = \frac{c(c+\ell_1\epsilon)}{c+\epsilon}\nu_1\mu_2 \tag{34} $$

$$ \nu = \frac{c(c+\epsilon)(c+\ell_1\epsilon)\mu_1\nu_1\nu_2}{(c+\epsilon)^2\mu_1+c(c+\ell_1\epsilon)\nu_1\nu_2} \tag{35} $$

$$ \ell = \ell_2+1. \tag{36} $$

Suppose that $\mu\geq\mu_1$. Then $\langle\mu,\nu,\ell\rangle$ is (c,ϵ)-achievable.

Proof We prove that a $\langle\mu,\nu,\ell\rangle$-step can be achieved by a slowdown, a $\langle\mu_2,\nu_2,\ell_2\rangle$-step, a speedup, and a $\langle\mu_1,\nu_1\ell_1\rangle$-step. Suppose $e\leq b\leq d$ and

$(c+\epsilon)\mu b = \nu(d+\ell e)$, and let $\Xi = {}^a\exists^b\mathrm{DTISP}(n^d, n^e)$. We must show that $\Xi \Vdash {}^a\exists^{\mu b}\mathrm{DTISP}(n^{c\mu b}, n^{\epsilon\mu b})$. Let

$$x \;=\; \mu b/\mu_1 \;=\; \frac{c(c+\ell_1\epsilon)\nu_1\mu_2}{(c+\epsilon)\mu_1}b. \tag{37}$$

We have $x \geq b \geq e$ since $\mu \geq \mu_1$. If $x \leq d$, a speedup inference will give

$$\Xi \vdash {}^a\exists^x\forall^b\mathrm{DTISP}(n^{d-x+e}, n^e). \tag{38}$$

Assume for the moment that $b \leq d - x + e$, and thus certainly $x \leq d$. We show that a $\langle \mu_2, \nu_2, \ell_2\rangle$ step can be applied to (38). We have

$$\begin{aligned} d + (\ell_2+1)e \;&=\; d + \ell e \;=\; \frac{(c+\epsilon)\mu b}{\nu} \\ &=\; c(c+\ell_1\epsilon)\nu_1\mu_2\left(\frac{1}{(c+\epsilon)\mu_1} + \frac{c+\epsilon}{c(c+\ell_1\epsilon)\nu_1\nu_2}\right)b \\ &=\; \left(\frac{c(c+\ell_1\epsilon)\nu_1\mu_2}{(c+\epsilon)\mu_1} + \frac{(c+\epsilon)\mu_2}{\nu_2}\right)b \;=\; x + \frac{(c+\epsilon)\mu_2}{\nu_2}b, \end{aligned}$$

whence

$$(c+\epsilon)\mu_2 b \;=\; \nu_2((d-x+e)+\ell_2 e).$$

Thus, by (38) and the (c,ϵ)-achievability of $\langle \mu_2, \nu_2, \ell_2\rangle$,

$$\Xi \;\Vdash\; {}^a\exists^x\forall^{\mu_2 b}\mathrm{DTISP}(n^{c\mu_2 b}, n^{\epsilon\mu_2 b}). \tag{39}$$

On the other hand, if $b > d - x + e$, we apply a speedup inference to Ξ with parameter $d - b + e$, followed by a weakening to obtain:

$$\begin{aligned} \Xi \;&\vdash\; {}^a\exists^{d-b+e}\forall^b\mathrm{DTISP}(n^b, n^e) \\ &\vdash^{w}\; {}^a\exists^x\forall^{\mu_2 b}\mathrm{DTISP}(n^{c\mu_2 b}, n^e) \end{aligned} \tag{40}$$

since $\mu_2 \geq 1$. This is the same as (39) but with a larger space bound. (The larger space bound is harmless, as the next inference will be a slowdown.)

The argument splits into two cases again, now depending on whether $x \leq c\mu_2 b$ or not. First consider the case $x \leq c\mu_2 b$. This certainly holds if $\langle \mu_1, \nu_1, \ell_1\rangle$ is useful, since then $(c+\ell_1\epsilon)\nu_1/(c+\epsilon) < 1$. In this case, a slowdown inference, applied to (39) or (40), gives

$$\Xi \;\Vdash\; {}^a\exists^x\mathrm{DTISP}(n^{c^2\mu_2 b}, n^{\epsilon c\mu_2 b}). \tag{41}$$

Since (37) gives

$$(c+\epsilon)\mu_1 x = \nu_1(c^2\mu_2 b + \ell_1\epsilon c\mu_2 b),$$

a $\langle \mu_1, \nu_1, \ell_1 \rangle$ step applied to (41) yields

$$\Xi \Vdash {}^a\exists^{\mu_1 x}\mathrm{DTISP}(n^{c\mu_1 x}, n^{\epsilon\mu_1 x}) = {}^a\exists^{\mu b}\mathrm{DTISP}(n^{c\mu b}, n^{\epsilon\mu b}).$$

Now consider the case $x \geq c\mu_2 b$. Picking up from (39) or (40), a slowdown and a weakening give

$$\begin{array}{rcl} \Xi & \Vdash & {}^a\exists^{x}\mathrm{DTISP}(n^{cx}, n^{\epsilon x}) \\ & \stackrel{w}{\vdash} & {}^a\exists^{\mu_1 x}\mathrm{DTISP}(n^{c\mu_1 x}, n^{\epsilon\mu_1 x}) = {}^a\exists^{\mu b}\mathrm{DTISP}(n^{c\mu b}, n^{\epsilon\mu b}) \end{array}$$

This proves Lemma 31. □

Lemma 31 requires that $\mu \geq \mu_1$. The case where this does not hold can be handled a method analogous to the "max" method of Lemma 12.

Lemma 32 *Let $\langle \mu_1, \nu_1, \ell_1 \rangle$ and $\langle \mu_2, \nu_2, \ell_2 \rangle$ be (c, ϵ)-achievable. Set*

$$\mu = \max\left\{ \frac{c(c+\ell_1\epsilon)}{c+\epsilon}\nu_1\mu_2, \mu_1 \right\}. \tag{42}$$

$$\nu = \frac{c(c+\epsilon)(c+\ell_1\epsilon)\mu_1\nu_1\nu_2}{(c+\epsilon)^2\mu_1 + c(c+\ell_1\epsilon)\nu_1\nu_2} \tag{43}$$

$$\ell = \ell_2 + 1. \tag{44}$$

Then $\langle \mu, \nu, \ell \rangle$ is (c, ϵ)-achievable.

Proof If $\mu_1 \leq c(c+\ell_1\epsilon)\nu_1\mu_2/(c+\epsilon)$, the previous lemma implies the result. Otherwise, set $\mu_2' = (c+\epsilon)\mu_1/(c(c+\ell_1\epsilon)\nu_1) > \mu_2$. By Lemma 29, $\langle \mu_2', \nu_2, \ell_2 \rangle$ is (c, ϵ)-achievable. Lemma 31 applied to the triples $\langle \mu_1, \nu_1, \ell_1 \rangle$ and $\langle \mu_2', \nu_2, \ell_2 \rangle$ gives the desired result. □

The constructions of the previous two lemmas show how to compose two (c, ϵ)-achievable triples to form a third one. It is again helpful to think of the triple $\langle \mu_1, \nu_1, \ell_1 \rangle$ as transforming the triple $\langle \mu_2, \nu_2, \ell_2 \rangle$. The mapping notation

$$\langle \mu_1, \nu_1, \ell_1 \rangle : \langle \mu_2, \nu_2, \ell_2 \rangle \mapsto \langle \mu, \nu, \ell \rangle$$

is used to indicate that equations (34)-(36) hold. Similarly,

$$\langle \mu_1, \nu_1, \ell_1 \rangle : \langle \mu_2, \nu_2, \ell_2 \rangle \mapsto^{\max} \langle \mu, \nu, \ell \rangle$$

indicates that equations (42)-(44) hold.

To better understand the action of these transformations, we can rewrite equations (34) and (35) as

$$\frac{1}{\mu} = \frac{c+\epsilon}{c(c+\ell_1\epsilon)\nu_1} \cdot \frac{1}{\mu_2} \tag{45}$$

$$\frac{1}{\nu} = \frac{1}{\tau(\mu_1,\nu_1,\ell_1)} - \frac{c+\epsilon}{c(c+\ell_1\epsilon)\nu_1}\left(\frac{1}{\tau(\mu_1,\nu_1,\ell_1)} - \frac{1}{\nu_2}\right), \tag{46}$$

where

$$\tau(\mu_1,\nu_1,\ell_1) = (c+\epsilon)\mu_1\left(1 - \frac{c+\epsilon}{c(c+\ell_1\epsilon)\nu_1}\right).$$

To write these equations more compactly, recall that

$$\rho(\mu_1,\nu_1,\ell_1) = \frac{c(c+\ell_1\epsilon)\nu_1}{c+\epsilon}, \tag{47}$$

and set $R_1 = \rho(\mu_1,\nu_1,\ell_1)$ and $T_1 = \tau(\mu_1,\nu_1,\ell_1)$. Then

$$T_1 = (c+\epsilon)\mu_1(1 - 1/R_1),$$

and equations (45) and (46) become

$$\frac{1}{\mu} = \frac{1}{R_1} \cdot \frac{1}{\mu_2} \quad \text{and} \quad \frac{1}{\nu} = \frac{1}{T_1} - \frac{1}{R_1}\left(\frac{1}{T_1} - \frac{1}{\nu_2}\right). \tag{48}$$

Equations (46) and (48) show the action of $\langle \mu_1, \nu_1, \ell_1 \rangle$ on the triple $\langle \mu_2, \nu_2, \ell_2 \rangle$ is in effect defining the value of $1/\nu$ by contracting $1/\nu_2$ towards $1/T_1$.

Another suggestive, and highly useful, way to rewrite Equations (35) and (46) is as

$$\frac{1}{\nu} = \frac{1}{(c+\epsilon)\mu_1} + \frac{1}{R_1\nu_2}. \tag{49}$$

The above constructions generalize those of Section 3.3. In particular, the $\epsilon = 0$ case is the same as the earlier results. In fact, most of our just obtained results can be obtained from those of Section 3.3 by replacing "ν_1" uniformly with "$(c+\ell_1\epsilon)\nu_1/(c+\epsilon)$". The extra complication, however, is the presence of the parameter ℓ: this makes the next section's analysis of alternation trading refutations substantially more difficult.

8 The refutations for DTISP lower bounds

We describe next the alternation trading refutations that give time/space (DTISP) lower bounds for algorithms for SAT. Perhaps surprisingly, the

refutations for DTISP lower bounds do *not* follow the pattern of refutations that were used in Section 5 for the lower bounds for DTS ($\epsilon = 0$) algorithms. Namely, the proof of Theorem 24 showed that if there is a refutation (in the DTS setting) if and only if it can be obtained by letting $\langle \mu_1, \nu_1 \rangle = \langle 1, c-1 \rangle$ and defining $\langle \mu_{i+1}, \nu_{i+1} \rangle$ by

$$\langle \mu_1, \nu_1 \rangle : \langle \mu_i, \nu_i \rangle \mapsto \langle \mu_{i+1}, \nu_{i+1} \rangle,$$

and finally obtaining some $\nu_i < 1/c$. We initially conjectured that we could use the same methods for DTISP refutations: Let $\langle \mu_1, \nu_1, \ell_1 \rangle$ be $\langle 1, c+\epsilon-1, 1 \rangle$, and define $\langle \mu_{i+1}, \nu_{i+1}, \ell_{i+1} \rangle$ by

$$\langle \mu_1, \nu_1, \ell_1 \rangle : \langle \mu_i, \nu_i, \ell_i \rangle \mapsto \langle \mu_{i+1}, \nu_{i+1}, \ell_{i+1} \rangle.$$

Our hope was that this would always suffice to obtain alternation trading refutations. This turned out, in computer experiments, to be false; namely, there are values c and ϵ which have refutations, but for which no $i > 0$ has $\rho(\mu_i, \nu_i, \ell_i) < 1$.

This greatly complicates a computer-based search for alternation trading refutations. Fix values for c and ϵ, and consider trying out all possible constructions of (c, ϵ)-achievable triples. It turns out that a blind, depth-first search for refutations based on the constructions (A)-(C) defined below will yield an immense set of achievable triples, even after discarding subsumed triples. This kind of blind search can eventually find a refutation if one exists; however, if no refutation exists, the process never stops and yields no information about the existence of a refutation.

In order to overcome this, we shall define two expanded notions of "subsumed" which will allow the computer-based search to more aggressively prune (c, ϵ)-achievable triples from the search space. This will be completely successful in reducing the size of the search space. Even more importantly, when there is no alternation trading refutation, the broader notions of subsumption allow the computer-based search to terminate quickly with a proof that no refutation exists (modulo round-off errors in the calculations).

The operations (A)-(E) for introducing (c, ϵ)-achievable triples are:

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

(B) Suppose $\langle \mu_1, \nu_1, \ell_1 \rangle$ and $\langle \mu_2, \nu_2, \ell_2 \rangle$ are (c, ϵ)-achievable and that $\mu_1 \leq c(c + \ell_1\epsilon)\nu_1\mu_2/(c + \epsilon)$. Then $\langle \mu, \nu, \ell \rangle$ is (c, ϵ)-achievable, where

$$\langle \mu_1, \nu_1, \ell_1 \rangle : \langle \mu_2, \nu_2, \ell_2 \rangle \mapsto \langle \mu, \nu, \ell \rangle.$$

(C) Suppose $\langle \mu_1, \nu_1, \ell_1 \rangle$ and $\langle \mu_2, \nu_2, \ell_2 \rangle$ are (c, ϵ)-achievable and that $\mu_1 > c(c + \ell_1\epsilon)\nu_1\mu_2/(c + \epsilon)$. Then $\langle \mu, \nu \rangle$ is (c, ϵ)-achievable, where

$$\langle \mu_1, \nu_1, \ell_1 \rangle : \langle \mu_2, \nu_2, \ell_2 \rangle \mapsto^{\max} \langle \mu, \nu, \ell \rangle.$$

(D) If $\langle \mu_2, \nu_2, \ell_2 \rangle$ is (c, ϵ)-achievable, then so is $\langle \mu, \nu, \ell \rangle$, where

$$\langle 1, 1, 1 \rangle : \langle \mu_2, \nu_2, \ell_2 \rangle \mapsto \langle \mu, \nu, \ell \rangle.$$

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

As before, the ABE-triples are defined to be those triples that can be inferred by operations (A), (B), and (E). Analogously to the earlier results for DTS-refutations, Theorem 49 states that, for any fixed values of c and ϵ, there is a DTISP-refutation if and only if some ABE-triple $\langle \mu, \nu, \ell \rangle$ has $\rho(\mu, \nu, \ell) < 1$ and $c\mu\epsilon < 1$.[1]

Our computer-based search focuses on finding an ABC-triple with ρ-value < 1. In light of Lemma 27, it is also required that the triple satisfy $c\mu\epsilon < 1$. More generally, we shall require that $c\mu\epsilon \leq R$ holds for all triples generated during the computer search, where $R = \rho(\mu, \nu, \ell)$. The reason for this is that our theoretical results below depend on having $c\mu\epsilon \leq R$. Our experimental results are that $c\mu\epsilon \leq R$ does in fact always hold, even though we have been unable to prove the condition should hold.

The computer search proceeds in rounds, or "stages". At each stage, there is a set Γ of derived (c, ϵ)-achievable triples. Initially, Γ contains just $\langle 1, c+\epsilon-1, 1 \rangle$. At each stage, two triples τ_1 and τ_2 are chosen from Γ, and a new triple τ is obtained by operation (B) or (C). It is checked that the new triple has $c\mu\epsilon \leq R$, and then the triple will be either pruned (discarded), or is added to Γ. The process terminates either when there is no new triple available to add to Γ or when a triple with ρ value < 1 is generated.

New triples are generated with operations (B) and (C), instead of with operations (B) and (E), essentially because the only point of using the subsumption operation (E) is to weaken a triple (as in the proof of Lemma 32 where μ_2 was increased to the value μ_2') in order that it may be used in operation (B). More precisely, it is easy to prove the following lemma by induction on the number of applications of operations (B) and (E).

[1] The proof of Theorem 49 is postponed until Section 9 as it is rather similar to the earlier proofs for DTS refutations.

Lemma 33 *If a triple τ can be derived from Γ by* (B) *and* (E) *operations, then there is a triple τ' derivable from Γ using only* (B) *and* (C) *operations, such that τ' subsumes τ. Furthermore the minimum number of* (B) *and* (C) *operations needed to derive τ' is at most the number of* (B) *operations used in the derivation of τ.*

Notation To streamline the discussion, we denote a triple $\langle \mu, \nu, \ell \rangle$ by the Greek letter τ; and we let $R = \rho(\mu, \nu, \ell)$. If there are subscripts, superscripts, or other marks on τ, they are applied uniformly to its associated values. For example, $\tau_i = \langle \mu_i, \nu_i, \ell_i \rangle$ and has ρ value R_i.

We write $\tau_1[\tau_2]$ to denote the triple τ obtained by operation (B).

Although the computer search uses (B) and (C) operations, our theoretical analysis below is based on (B) and (E) derivations. This is justified by Lemma 33.

Definition A derivation of a triple from Γ is viewed as a tree. The leaves of the tree are members of Γ, and internal nodes are labeled with triples that are inferred by (B) or (E) operations from their children. The *height* of a derivation is the maximum number of (B) operations along any branch of the tree.

A *contradiction* from Γ is a triple τ derivable from Γ that has $R < 1$.

Definition Let Γ be a set of triples and τ a (new) triple. We say τ may be *pruned* provided that any derivation of a contradiction from $\Gamma \cup \tau$ has height greater than or equal to the minimum height of a derivation of a contradiction from Γ.

For example, it is clear that if τ is subsumed by some member of Γ, then τ may be pruned. The next definition gives a less trivial, and more useful, notion of subsumption.

Definition The triple τ_1 *R-subsumes* τ_2 provided that $\mu_1 \leq \mu_2$ and $R_1 \leq R_2$ and $\ell_1 \leq \ell_2$.

By the definition of $\rho(\mu, \nu, \ell)$, if τ_1 subsumes τ_2, then certainly τ_1 R-subsumes τ_2. Thus R-subsumption is a more general notion than subsumption. In addition, we claim that if τ is R-subsumed by some triple in Γ, then τ may be pruned from Γ. This is proved by induction on the height of derivations from $\Gamma \cup \{\tau\}$ that contain (B) and (E) operations, using the following lemma for the key induction argument.

Lemma 34 *Suppose τ_0 and τ_1 R-subsume τ_3 and τ_4, respectively. In addition, suppose $R_3 \geq 1$. Let $\tau_2 = \tau_0[\tau_1]$ and $\tau_5 = \tau_3[\tau_4]$. Then either τ_2 R-subsumes τ_5, or τ_4 R-subsumes τ_5.*

Proof We have $\mu_2 = R_0\mu_1 \leq R_3\mu_4 = \mu_5$, so $\mu_2, \mu_4 \leq \mu_5$. Also, $\ell_2 = \ell_1 + 1 \leq \ell_4 + 1 = \ell_5$. It thus suffices to show that either $R_2 \leq R_5$ or $R_4 \leq R_5$. Referring back to equations (47) and (49) for R and $1/\nu$, we have

$$\begin{aligned}
\frac{1}{R_5} &= \frac{c+\epsilon}{c(c+\ell_5\epsilon)}\left(\frac{1}{(c+\epsilon)\mu_3} + \frac{1}{R_3}\frac{c(c+\ell_4\epsilon)}{c+\epsilon}\frac{1}{R_4}\right) \\
&= \frac{1}{c(c+\epsilon\ell_4+\epsilon)}\frac{1}{\mu_3} + \left(1 - \frac{\epsilon}{c+\epsilon\ell_4+\epsilon}\right)\frac{1}{R_3}\frac{1}{R_4} \\
&= \frac{1}{c(c+\epsilon\ell_4+\epsilon)}\left(\frac{1}{\mu_3} - c\epsilon\frac{1}{R_3}\frac{1}{R_4}\right) + \frac{1}{R_3}\frac{1}{R_4} \qquad (50)
\end{aligned}$$

with analogous equations holding for $1/R_2$ (replacing subscripts 3,4,5 with 0,1,2, respectively).

Suppose the quantity in parentheses in the last equation is negative, so $\left(\frac{1}{\mu_3} - c\epsilon\frac{1}{R_3}\frac{1}{R_4}\right) < 0$. Then

$$\frac{1}{R_5} \;\leq\; \frac{1}{R_3}\frac{1}{R_4} \;\leq\; \frac{1}{R_4}$$

since $R_3 \geq 1$. It follows that $R_4 \leq R_5$, so τ_4 R-subsumes τ_5.

Otherwise, that quantity is non-negative. Then,

$$\begin{aligned}
\frac{1}{R_5} &\leq \frac{1}{c(c+\epsilon\ell_1+\epsilon)}\left(\frac{1}{\mu_3} - c\epsilon\frac{1}{R_3}\frac{1}{R_4}\right) + \frac{1}{R_3}\frac{1}{R_4} \\
&= \frac{1}{c(c+\epsilon\ell_1+\epsilon)}\frac{1}{\mu_3} + \left(1 - \frac{\epsilon}{c+\epsilon\ell_1+\epsilon}\right)\frac{1}{R_3}\frac{1}{R_4} \\
&\leq \frac{1}{c(c+\epsilon\ell_1+\epsilon)}\frac{1}{\mu_0} + \left(1 - \frac{\epsilon}{c+\epsilon\ell_1+\epsilon}\right)\frac{1}{R_0}\frac{1}{R_1} \;=\; \frac{1}{R_2}.
\end{aligned}$$

The first inequality above follows from $\ell_1 \leq \ell_4$; the second inequality from $\mu_0 \leq \mu_3$, $R_0 \leq R_3$, and $R_1 \leq R_4$. Thus $R_5 \geq R_2$, so τ_2 R-subsumes τ_5. □

Corollary 35 *If τ is R-subsumed by a triple in Γ, then τ may be pruned.*

Corollary 35 is proved using induction on the height of derivations. Namely, any triple derivable from $\Gamma \cup \tau$ with a derivation of height h is R-subsumed by some triple derivable from just Γ with a derivation of height $\leq h$. We leave the details to the reader.

Note that the requirement in Lemma 34 that $R_3 \geq 1$ is harmless, since if any of R_0, R_1, R_3, R_4 are < 1, then a contradiction has already been reached.

The ability to prune R-subsumed triples reduces the search space considerably, but it still leads to large search spaces and, when no contradiction exists, even to unbounded searches. Accordingly, we next define a yet-stronger form of subsumption, called "dual subsumption": it requires two triples to "dual subsume" a new triple.

For the intuition of what it means for τ_1 and τ_3 to "dual-subsume" τ_2, recall from equation (49) that if $\tau' = \tau_i[\tau]$, then ν' is calculated as

$$\frac{1}{\nu'} \;=\; \frac{1}{(c+\epsilon)\mu_i} + \frac{1}{R_i}\frac{1}{\nu} \tag{51}$$

The definition of dual-subsumption is based on the effectiveness of each τ_i at producing a small value for ν'. Letting L_i be the line with y-intercept $1/((c+\epsilon)\mu_i)$ and slope $1/R_i$, then the effectiveness of τ_i is represented by the height of the line L_i — the higher the line L_i is, the better τ_i is. This is illustrated in Figure 1, where the two lines L_1 and L_3 jointly dominate L_3 in the sense that L_2 lies below the maximum of the lines L_1 and L_3 for all values of $1/\nu$. This will be enforced by having $v_{23} \leq v_{13} \leq v_{12}$, where

$$v_{ij} \;=\; \frac{\frac{1}{(c+\epsilon)\mu_i} - \frac{1}{(c+\epsilon)\mu_j}}{\frac{1}{R_j} - \frac{1}{R_i}}$$

is the intersection of lines L_i and L_j.

Definition The triples τ_1 and τ_3 *dual-subsume* τ_2 provided the following four conditions hold.

(ds1) $\frac{1}{\mu_1} > \frac{1}{\mu_2} > \frac{1}{\mu_3}$,

(ds2) $\frac{1}{R_1} < \frac{1}{R_2} < \frac{1}{R_3}$,

(ds3) $\ell_1 \leq \ell_2$ and $\ell_3 \leq \ell_2$,

and

$$\frac{\frac{1}{\mu_2} - \frac{1}{\mu_3}}{\frac{1}{R_3} - \frac{1}{R_2}} \;\leq\; \frac{\frac{1}{\mu_1} - \frac{1}{\mu_3}}{\frac{1}{R_3} - \frac{1}{R_1}} \;\leq\; \frac{\frac{1}{\mu_1} - \frac{1}{\mu_2}}{\frac{1}{R_2} - \frac{1}{R_1}}. \tag{52}$$

We write "(52^{rv})" to denote the last line with the directions of the inequalities reversed.

The notions of "R-subsume" and "dual-subsume" can be generalized to "multisubsume" as follows. Let τ be a triple and Γ be a finite set of triples

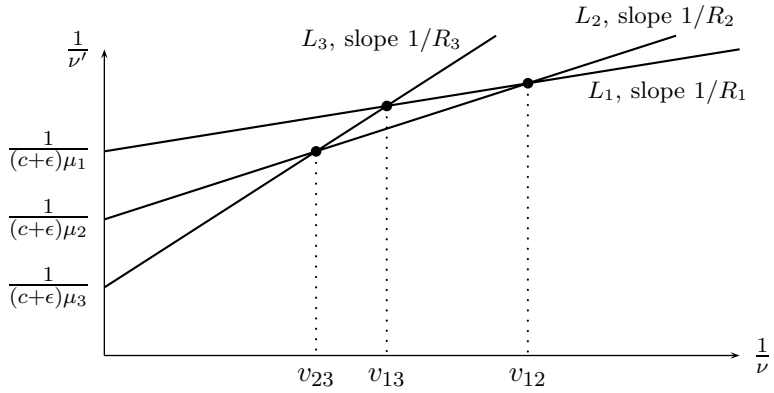


Figure 1: L_1 and L_3 dominate L_2.

such that (1) for each $\tau_i \in \Gamma$, $\ell_i \leq \ell$, and (2) for positive values of $1/\nu$, the line L associated with τ is bounded above by the maximums of the lines L_i associated with τ_i's in Γ. Another way to state condition (2) is that the infinite convex polytope in the first quadrant which is bounded below by the lines L_i is above the line L.

Lemma 36 *Γ multisubsumes τ if and only if there are τ_i and τ_j in Γ which dual-subsume τ or there is a τ_i in Γ that R-subsumes τ.*

The proof of the lemma is trivial: Consider a closest vertex of the polytope to the line L. Then either the two triples corresponding to the edges adjacent to this vertex dual-subsume L, or one of them R-subsumes L.

This immediately implies a transitivity property for dual- and multi-subsumption:

Corollary 37 *Suppose that τ_1 and τ_2 dual-subsume τ_5, that τ_3 and τ_4 dual-subsume τ_6, and that τ_5 and τ_6 dual-subsume τ_7. Then, some two of τ_1, τ_2, τ_3 and τ_4 dual-subsume τ_7. More generally, if Γ multisubsumes each triple in Δ, and Δ multisubsumes τ, then Γ multisubsumes τ.*

We would like to extend Corollary 35 (which stated that R-subsumed triples may be discarded) to conclude that dual-subsumed triples may be discarded during the computer search for contradictions. We need an additional assumption however, namely that all generated triples satisfy the property that $c\mu\epsilon \leq R$. Theorem 44 will state this property precisely, but first we prove a series of preliminary results as Lemmas 38-43.

Lemma 38 *The validity of the inequalities in* (52) *is unchanged by any of the following changes to the values of μ_i or R_i:*

(a) *Multiplying each value* $\frac{1}{\mu_i}$ *by a scalar* $\alpha > 0$.

(b) *Multiplying each value* $\frac{1}{R_i}$ *by a scalar* $\alpha > 0$.

(c) *Adding a constant* α *to each value* $\frac{1}{\mu_i}$.

(d) *Adding a constant* α *to each value* $\frac{1}{R_i}$.

(e) *Replacing each* $\frac{1}{\mu_i}$ *with* $\frac{1}{\mu_i} + \alpha\frac{1}{R_i}$ *where* α *is a scalar.*

(f) *Replacing each* $\frac{1}{R_i}$ *with* $\frac{1}{R_i} + \alpha\frac{1}{\mu_i}$ *where* α *is a scalar.*

Furthermore, (g) *if* (52) *holds then after swapping the values of* $\frac{1}{\mu_i}$ *with* $\frac{1}{R_i}$, *property* (52^{rv}) *holds, and vice-versa.*

The proof of Lemma 38 is trivial. The case (f) can be proved using (g), then (e), and then (g) again.

Lemma 39 *Let* τ_1 *and* τ_3 *dual-subsume* τ_2 *with* $R_3 \geq 1$. *Let* τ *be a triple that satisfies* $c\mu\epsilon \leq R$. *Set* $\tau_i' = \tau[\tau_i]$ *for all* i. *Then* τ_1' *and* τ_3' *multisubsume* τ_2'.

Proof The definition of the triples τ_i' gives

$$\frac{1}{\mu_i'} = \frac{1}{R}\frac{1}{\mu_i} \tag{53}$$

$$\frac{1}{R_i'} = \frac{1}{c(c+\epsilon\ell_i+\epsilon)}\left(\frac{1}{\mu} - c\epsilon\frac{1}{R}\frac{1}{R_i}\right) + \frac{1}{R}\frac{1}{R_i}., \tag{54}$$

similarly to equation (50). Note that the quantity in the parentheses must be nonnegative, since each $R_i \geq 1$ and since $c\mu\epsilon \leq R$. Consequently, R_i' will increase if the value of ℓ_i is increased.

We claim that we may assume w.l.o.g. that $\ell_1 = \ell_2 = \ell_3$, and thus $\ell_1' = \ell_2' = \ell_3'$. To see this, consider increasing the values of ℓ_1 and ℓ_3 to equal ℓ_2 while keeping the values R_i fixed. (This will keep (ds3) satisfied.) To keep $R_i = (c(c+\ell_i)/(c+\epsilon))\nu_i$ fixed, it is necessary to also decrease the values of ν_1 and ν_3. These changes however, do not affect the hypothesis of dual subsumption. In addition, as just remarked, this only increases the values R_1' and R_3', and it leaves the values of μ_1' and μ_3' unchanged. This only makes it harder to establish the desired multisubsumption; indeed, it shifts the lines L_1' and L_3' downward while L_2' is remains unchanged. (L_i' is the line defined by μ_i' and R_i', similarly as in Figure 1.)

By (53) and (ds1),

(ds1)′ $\frac{1}{\mu_1'} > \frac{1}{\mu_2'} > \frac{1}{\mu_3'}$.

From (54) and the assumption that $\ell_1 = \ell_2 = \ell_3$, it follows that $1/R_i'$ is a linear function of $1/R_i$. Therefore,

(ds2)′ $\frac{1}{R_1'} < \frac{1}{R_2'} < \frac{1}{R_3'}$

For the same reasons, and by parts (a), (b) and (d) of Lemma 38, condition (52) holds for the triples τ_1', τ_2', and τ_3'. Note that for part (a), the scalar α is $1/R$ and is positive. For part (b), the scalar α is $(c+\epsilon\ell_i)/(R(c+\epsilon(\ell_i+1)))$; this is the same for all i and positive.

It follows that τ_1' and τ_3' dual-subsume, and hence multisubsume, τ_2'. □

Corollary 40 *Suppose τ_1 and τ_3 multisubsume τ_2, and $c\mu\epsilon \leq R$ and $1 \leq R_1, R_3, R$. Let $\tau_i' = \tau[\tau_i]$. Then the four triples $\tau_1, \tau_3, \tau_1', \tau_3'$ multisubsume τ_2'.*

Proof If τ_1 and τ_3 dual-subsume τ_2, the corollary follows from the previous lemma. Otherwise, one of τ_1 or τ_3 R-subsumes τ_2.

Suppose that τ_1 R-subsumes τ_2. Since τ R-subsumes itself, Lemma 34 implies that one of τ_1' or τ_2 R-subsumes τ_2'. If it is τ_2 that R-subsumes τ_2', then the transitivity of R-subsumption implies that τ_1 also R-subsumes τ_2'.

Likewise, if τ_3 R-subsumes τ_2, then either τ_3' or τ_3 R-subsumes τ_2. This suffices to prove the corollary. □

Lemma 41 *Let τ_1 and τ_3 dual-subsume τ_2. Let τ also be a triple, and let $\tau_i' = \tau_i[\tau]$. Then τ_3' and τ_1' multisubsume τ_2'.*

Proof Each μ_i' is equal to $R_i\mu$. Thus, from (ds2),

(ds1^{rv})′ $\frac{1}{\mu_3'} > \frac{1}{\mu_2'} > \frac{1}{\mu_1'}$.

Suppose $\frac{1}{\nu} \geq v_{23}$. Referring back to Figure 1, it is evident that $\frac{1}{\nu_3'} \geq \frac{1}{\nu_2'}$ since L_3 is above L_2 in this range. Hence, $\mu_3' \leq \mu_2'$, and $\nu_3' \leq \nu_2'$, and $\ell_3' = \ell_2'$, and therefore τ_3' subsumes (and hence R-subsumes and multisubsumes) the triple τ_2'.

On the other hand, suppose $\frac{1}{\nu} < v_{23}$. From Figure 1 again, we have $\frac{1}{\nu_1'} > \frac{1}{\nu_2'} > \frac{1}{\nu_3'}$. Since $\ell_1' = \ell_2' = \ell_3' = \ell + 1$, this implies

(ds2^{rv})′ $\frac{1}{R_3'} < \frac{1}{R_2'} < \frac{1}{R_1'}$,

We claim that τ_3' and τ_1' dual-subsume τ_2'. For this, we must show that τ_1', τ_2', and τ_3' satisfy the conditions of (52^{rv}). Since each $\ell_i' = \ell + 1$, we have

$$\begin{aligned} \frac{1}{R_i'} &= \frac{c+\epsilon}{c(c+\epsilon\ell+\epsilon)} \left(\frac{1}{c+\epsilon} \cdot \frac{1}{\mu_i} + \frac{1}{\nu} \cdot \frac{1}{R_i} \right) \\ \frac{1}{\mu_i'} &= \frac{1}{\mu} \cdot \frac{1}{R_i} \end{aligned}$$

Hence, using part (g) of Lemma 38, and then parts (a), (b), and (f), it follows that τ_1', τ_2', and τ_3' satisfy the conditions of (52^{rv}). □

Corollary 42 *Suppose τ_1 and τ_3 multisubsume τ_2 and that $1 \leq R_2$. Let $\tau_i' = \tau_i[\tau]$. Then the five triples $\tau, \tau_1, \tau_3, \tau_1', \tau_3'$ multisubsume τ_2'.*

Proof If τ_1 and τ_3 dual-subsume τ_2, the corollary follows from the previous lemma. Otherwise, one of τ_1 or τ_3 R-subsumes τ_2.

Suppose that τ_1 R-subsumes τ_2. Since τ R-subsumes itself, Lemma 34 implies that one of τ_1' or τ R-subsumes τ_2'. Likewise, if τ_3 R-subsumes τ_2, then either τ_3' or τ R-subsumes τ_2.

This suffices to prove the corollary. □

Definition $\Gamma[\Gamma]$ is the set of triples $\{\tau_1[\tau_3] : \tau_1, \tau_3 \in \Gamma\}$.

Lemma 43 *Suppose every triple in Γ satisfies $1 < R$ and $c\mu\epsilon \leq R$. Also suppose Γ multisubsumes τ_2' and τ_2'', and that $1 \leq R_2'$. Let $\tau_2 = \tau_2'[\tau_2'']$. Then $\Gamma \cup \Gamma[\Gamma]$ multisubsumes τ_2.*

Proof By assumption, there are $\tau_1', \tau_1'', \tau_3', \tau_3'' \in \Gamma$ so that τ_1' and τ_3' multisubsume τ_2', and τ_1'' and τ_3'' multisubsume τ_2''. From Corollary 40, we have

$$\tau_1'', \tau_3'', \tau_1'[\tau_1''], \tau_1'[\tau_3''] \quad \text{multisubsume} \quad \tau_1'[\tau_2''],$$

and

$$\tau_1'', \tau_3'', \tau_3'[\tau_1''], \tau_3'[\tau_3''] \quad \text{multisubsume} \quad \tau_3'[\tau_2''].$$

And, from Corollary 42,

$$\tau_2'', \tau_1', \tau_3', \tau_1'[\tau_2''], \tau_3'[\tau_2''] \quad \text{multisubume} \quad \tau_2'[\tau_2''].$$

Now, by transitivity of multisubsumption (Corollary 37),

$$\tau_1', \tau_3', \tau_1'', \tau_3'', \tau_1'[\tau_1''], \tau_1'[\tau_3''], \tau_3'[\tau_1''], \tau_3'[\tau_3''] \quad \text{multisubume} \quad \tau_2'[\tau_2''].$$

□

Theorem 44 *Suppose* Γ *is a set of triples which satisfy* $1 < R$ *and* $c\mu\epsilon < R$. *Further suppose that* Γ *multisubsumes every triple from* $\Gamma[\Gamma]$. *Then there is no contradiction derivable from* Γ.

Proof Let τ be a triple derivable from Γ. We claim that Γ multisubsumes τ. The claim is proved by induction on the number of steps in the derivation of τ from Γ. For the induction step, Lemma 43 implies that τ is multisubsumed by $\Gamma \cup \Gamma[\Gamma]$; thus by the transitivity of multisubsumption, τ is multisubsumed by Γ.

Finally, note that if τ is R-subsumed by τ_1 or is dual subsumed by τ_0 and τ_1, then $R \geq R_1$. If follows that $R > 1$ since $R_1 > 1$ for every τ_1 in Γ. □

We can now describe and justify the algorithm behind our computer-based search for alternation trading refutations. The program runs in stages, maintaining a set Γ of triples that satisfy $1 < R$ and $c\mu\epsilon \leq R$. Initially, Γ contains the single triple $\langle 1, c+\epsilon-1, 1\rangle$. The algorithm loops, repeatedly generating new triples using operations of type (B) and (C). For each new τ, it does the following:

(i) If $c\mu\epsilon > R$, the program aborts and fails to give an answer.

(ii) If $c\mu\epsilon \leq R < 1$, the program has found a contradiction, and the program halts and reports there that an alternation trading refutation exists. This implies that SAT is not in DTISP(n^c, n^ϵ).

(iii) Otherwise the triple τ is added to Γ for the next iteration of the loop.

If the algorithm ever reaches a stage where no new triple is added to Γ, then it halts and reports that there is no alternation trading refutation for these values of c and ϵ.[2]

Figure 2 tabulates and graphs, for various values of ϵ, the maximum value of c for which there exists an alternation trading proof (with five digits of accuracy). In each case, the computer-based search was able to *prove* there is no alternation trading proof for the next value of c.

[2]In actuality, our algorithm is implemented somewhat more efficiently than what was described. The program keeps track of the convex polytope discussed around Lemma 36. It also aggressively seeks for contradictions by first iteratively transforming the initial triple $\langle 1, c+\epsilon-1, 1\rangle$ with the optimal choices of triples corresponding to the edges of the convex polytope, before forming other triples. We developed these more sophisticated search strategies in order to explore the behavior of (c, ϵ)-achievable triples (and before discovering the concept of "dual-subsumption"). However, in the end, we conjecture that these sophisticated search techniques are not substantially more effective than other, more straightforward strategies.

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

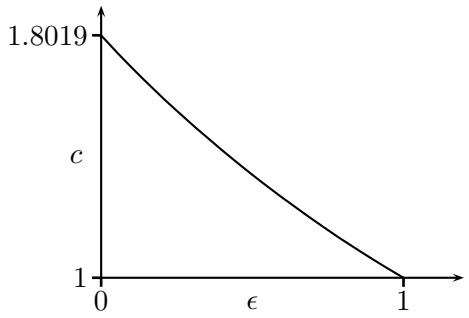


Figure 2: Showing the maximum value of c, as a function of ϵ, for which alternation trading proofs suffice to show that SAT is not in DTISP(n^c, n^ϵ).

The computer search for the values of c in Figure 2 was remarkably long in most cases. Indeed, the program often needed to run for tens of levels, and generate hundreds of triples. This is reported in Figure 3. For each value of ϵ and c, the table reports the number of rounds that were needed for the search, It also reports the total number of triples that were generated by the search and which could not be pruned immediately (and thus were retained for the next round of search). For instance, finding a refutation of $\epsilon = 0.5$ and $c = 1.3407$ required 44 rounds of the search, generating 406 intermediate triples. This means that the shortest refutation requires somewhere between 44 and 406 operations of type (B); this is a remarkably large number. Recall that each operation of type (B) can only be *approximated* by alternation trading derivations. Thus, the actual alternation trading refutations written out as R0$'_\epsilon$-R2$'_\epsilon$ steps will need to be much larger, well beyond the search capabilities of the Maple-based searches of [14].

It is remarkable that the computer-based search always succeeded to either find a refutation or prove that a refutation does not exist. We conjecture that this is guaranteed to happen: namely, either a refutation can be found by exhaustive search, or subsumption and dual-subsumption pruning will eventually establish that no refutation exists. However, our only evidence for the conjecture is the success of the computer-based searches in every case that we have tried.

It would be desirable to find a normal form for alternation trading proofs in the DTISP setting. However, even after looking carefully at the kinds of refutations generation by the computer searches, we have been unable to find a reasonable conjecture of a pattern that would allow direct (non-exhaustive)

ϵ	c	Number of Rounds	Number of Triples	Has Refutation
0.001	1.80084	7	167	No
	1.80083	11	455	Yes
0.01	1.79093	20	764	No
	1.79092	11	278	Yes
0.1	1.69619	248	3633	No
	1.69618	26	435	Yes
0.25	1.55243	249	2932	No
	1.55242	33	297	Yes
0.5	1.34071	203	1533	No
	1.34070	44	406	Yes
0.75	1.15766	155	1379	No
	1.15765	27	167	Yes
0.9	1.06012	146	454	No
	1.06011	19	88	Yes
0.99	1.00584	99	260	No
	1.00583	7	20	Yes
0.999	1.00059	3	3	No
	1.00058	24	10	Yes

Figure 3: Numbers of rounds and triples needed to find an alteration trading refutation or establish that none exists. A full table of results can be found online at
`http://math.ucsd.edu/ sbuss/ResearchWeb/npProofLimits/timespacebdsDS3.xls`.

search for refutations.

9 The limits of achievable DTISP constructions

This section establishes that refutations can do no better that what is possible using (c, ϵ)-achievable triples.[3] Furthermore, the only (c, ϵ)-achievable triples needed are those that are constructed using Lemmas 30-32.

Lemma 45 *Let Ξ be ${}^a\exists^b\mathrm{DTISP}(n^d, n^e)$, with $b \leq d$. Let $\mathcal{D}$ be a derivation that starts with the line Ξ and contains a non-empty* **(10)*** *pattern of inferences. Then $\mathcal{D}$ is subsumed by the (c, ϵ)-achievable triple $\langle 1, c{+}\epsilon{-}1, 1\rangle$.*

Proof We must show that the kind of refutation used in the proof of Lemma 30 is optimal. That proof considered speedup inferences with $x = b$; however, we now need to consider pairs of speedup-slowdown inferences with arbitrary values of x:

$$\begin{array}{rcl} \Xi & \vdash & {}^a\exists^{\max\{x,b\}}\forall^b\mathrm{DTISP}(n^{d-x+e}, n^e) \\ & \vdash & {}^a\exists^{\max\{x,b\}}\mathrm{DTISP}(n^{\max\{cx,cb,ce,c(d-x+e)\}}, n^{\cdots}). \end{array} \tag{55}$$

We wish to show that we can require $x = b$, at the expense of adding weakening inferences. This is argued similarly as in the proof of Lemma 15. First, if $x < b$, use a weakening inference to increase the value of d up to $d + b - x$, and then change the speedup to use $x = b$. Second, if $x > b$, use a weakening to increase the value of b to x before applying the speedup inference. In both cases, we get a new derivation with the same final line.

We can now assume $x = b$. Since $d \geq x = b \geq e$, the derivation (55) becomes

$$\begin{array}{rcl} \Xi & \vdash & {}^a\exists^b\forall^b\mathrm{DTISP}(n^{d-b+e}, n^e) \\ & \vdash & {}^a\exists^b\mathrm{DTISP}(n^{\max\{cb,c(d-b+e)\}}, n^{\max\{\epsilon b,\epsilon(d-b+e)\}}) \\ & = & {}^a\exists^b\mathrm{DTISP}(n^{d'}, n^{e'}). \end{array}$$

The derivation $\mathcal{D}$ can thus be assumed to consist of the following operations: (a) weakenings that increase d, (b) weakenings that increase b, and (c) speedup/slowdown pairs of the type just displayed. There will be at least one operation of type (c), and it has the effect of setting d' and e' so that

$$d' + e' \;=\; \max\{(c+\epsilon)b, (c+\epsilon)((d+e)-b)\}.$$

[3] Section 9 can be read independently of Section 8, with the sole exception of needing the definitions of operations (A)-(E) as given in Section 8.

(Refer to the proof of Lemma 30.) The same reasoning as used in Lemma 15 now shows that the derivation $\mathcal{D}$ is subsumed by $\langle 1, c{+}\epsilon{-}1, 1\rangle$.

Next we prove the central tool needed for showing derivations are subsumed by (c, ϵ)-achievable triples. This is a direct generalization of Lemma 17.

Lemma 46 *Suppose that A and B are balanced $\mathbf{0}/\mathbf{1}$-annotations, and that a derivation $\mathcal{D}$ has the inference pattern $\mathbf{1}A\mathbf{0}B$, and the first line Ξ of $\mathcal{D}$ has the form $\cdots {}^a\exists^b\mathrm{DTISP}(n^d, n^e)$ with $e \leq b \leq d$. Further suppose the subderivation corresponding to A is weakly subsumed by $\langle \mu_2, \nu_2, \ell_2\rangle$, and the subderivation corresponding to B is non-empty and subsumed (respectively, weakly subsumed) by $\langle \mu_1, \nu_1, \ell_1\rangle$. Then the entire derivation $\mathcal{D}$ is subsumed (respectively, weakly subsumed) by a triple $\langle \mu, \nu, \ell\rangle$ such that either*

$$\langle \mu_1, \nu_1, \ell_1\rangle : \langle \mu_2, \nu_2, \ell_2\rangle \mapsto^{\max} \langle \mu, \nu, \ell\rangle, \tag{56}$$

or

$$\langle 1, 1, 1\rangle : \langle \mu_2, \nu_2, \ell_2\rangle \mapsto \langle \mu, \nu, \ell\rangle. \tag{57}$$

If B is empty, then the derivation $\mathcal{D}$ is weakly subsumed by the $\langle \mu, \nu, \ell\rangle$ given by (57).

The triple defined by (57) is equal to

$$\langle \mu, \nu, \ell\rangle \;:=\; \langle c\mu_2,\; \frac{c(c+\epsilon)\nu_2}{(c+\epsilon)+c\nu_2},\; \ell_2 + 1\rangle.$$

Proof The derivation $\mathcal{D}$ has first line Ξ as above and its last line has the form $\cdots {}^a\exists^{x'}\mathrm{DTISP}(n^{u'}, n^{v'})$. We henceforth suppress mention of the "$\cdots$" prefix as it remains the same throughout the balanced derivation.

The first inference of $\mathcal{D}$ is a speedup,

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

We claim that w.l.o.g. $x \geq b$. Otherwise, arguing as before, we can insert a weakening inference to increase the value of d to $d + b - x$ and then do a speedup with $x = b$ to derive the same result. We thus henceforth assume $x \geq b$.

The $\mathbf{1}A\mathbf{0}$ part of $\mathcal{D}$ consists of a speedup, then a subderivation weakly subsumed by $\langle \mu_2, \nu_2, \ell_2 \rangle$, then a slowdown:

$$\begin{array}{rcll} {}^a\exists^b\mathrm{DTISP}(n^d, n^e) & \vdash & {}^a\exists^x\forall^b\mathrm{DTISP}(n^{d-x+e}, n^e) & \text{- by speedup} \\ & \vdots & \vdots & \text{(weakly subsumed by } \langle \mu_2, \nu_2, \ell_2 \rangle) \\ & \vdash & {}^a\exists^x\forall^y\mathrm{DTISP}(n^z, n^w) & \\ & \vdash & {}^a\exists^x\mathrm{DTISP}(n^u, n^v) & \text{- by slowdown} \end{array} \tag{58}$$

where $u = \max\{cx, cy, cz\}$ and $v = \max\{\epsilon x, \epsilon y, \epsilon z\}$, and, by the weak subsumption

$$z \;\geq\; \max\{c\mu_2 b, \frac{c}{c+\epsilon}\nu_2(d - x + (\ell_2 + 1)e)\}. \tag{59}$$

Suppose the B part of $\mathcal{D}$ is empty, so (58) is the last line of $\mathcal{D}$. We have

$$u \;\geq\; cz \;\geq\; c(c\mu_2)b$$

and

$$u \;\geq\; \max\{cx, c\frac{c}{c+\epsilon}\nu_2(d - x + (\ell_2 + 1)e)\}.$$

The right hand side of the latter inequality is minimized when

$$x \;=\; \frac{c\nu_2(d + (\ell_2 + 1)e)}{(c+\epsilon) + c\nu_2};$$

hence,

$$u \;\geq\; \frac{c}{c+\epsilon} \cdot \frac{c(c+\epsilon)\nu_2}{(c+\epsilon) + c\nu_2}(d + (\ell_2 + 1)e).$$

It follows that the derivation $\mathcal{D}$ is weakly subsumed by $\langle \mu, \nu, \ell \rangle$ as defined by (57).

Now assume B is non-empty. Referring back to (59), we claim that, w.l.o.g., $c\mu_2 b \leq \frac{c}{c+\epsilon}\nu_2(d - x + (\ell_2 + 1)e)$. If this does not hold, then we can increase the value of d to

$$d' \;=\; (x - (\ell_2 + 1)e) + (c+\epsilon)\frac{\mu_2}{\nu_2}b;$$

that is to say, we replace the $\mathbf{1}A\mathbf{0}$ part of $\mathcal{D}$ with

$$\begin{array}{rcll} \Xi & \stackrel{w}{\vdash} & {}^a\exists^b\mathrm{DTISP}(n^{d'}, n^e) & \text{- weakening} \\ & \vdash & {}^a\exists^x\forall^b\mathrm{DTISP}(n^{d'-x+e}, n^e) & \text{- by speedup} \\ & = & {}^a\exists^x\forall^b\mathrm{DTISP}(n^{(c+\epsilon)(\mu_2/\nu_2)b-(\ell_2+1)e+e}, n^e) & \\ & \Vdash & {}^a\exists^x\forall^{\mu_2 b}\mathrm{DTISP}(n^{c\mu_2 b}, n^{\epsilon\mu_2 b}) & \text{- by a } \langle \mu_2, \nu_2, \ell_2 \rangle \text{ step} \\ & = & {}^a\exists^x\forall^y\mathrm{DTISP}(n^z, n^w) & \text{- where } y = \mu_2 b \text{ and } z = c\mu_2 b. \\ & \vdash & {}^a\exists^x\mathrm{DTISP}(n^u, n^v) & \text{- by slowdown} \end{array}$$

In particular, z still satisfies the inequality (59), which is the crucial property needed for the $\mathbf{1}A\mathbf{0}$ portion of $\mathcal{D}$. We therefore may assume that

$$b \;\leq\; x \;\leq\; (d + (\ell_2 + 1)e) - (c + \epsilon)\frac{\mu_2}{\nu_2}b, \tag{60}$$

and so

$$(c + \epsilon)\frac{\mu_2}{\nu_2}b \;\leq\; d - x + (\ell_2 + 1)e. \tag{61}$$

The B part of the derivation $\mathcal{D}$ derives

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

Since this is weakly subsumed by $\langle \mu_1, \nu_1, \ell_1 \rangle$,

$$\begin{aligned} u' &\geq \max\{c\mu_1 x, \frac{c}{c+\epsilon}\nu_1(u + \ell_1 v)\} \\ &= \max\{c\mu_1 x, \frac{c}{c+\epsilon}\nu_1\frac{c + \ell_1\epsilon}{c}u\} \\ &\geq \max\{c\mu_1 x, \frac{c^2(c + \ell_1\epsilon)}{(c+\epsilon)^2}\nu_1\nu_2(d - x + (\ell_2 + 1)e)\}. \end{aligned} \tag{62}$$

The last inequality follows from $u \geq cz$ and (59). If B is also (non-weakly) subsumed by the triple, then the same reasoning shows

$$x' \;\geq\; \max\{\mu_1 x, \frac{c(c + \ell_1\epsilon)}{(c+\epsilon)^2}\nu_1\nu_2(d - x + (\ell_2 + 1)e)\}.$$

Referring to (62), we claim that, w.l.o.g., either

(i) $x = b$ and $c\mu_1 x \geq \frac{c^2(c+\ell_1\epsilon)}{(c+\epsilon)^2}\nu_1\nu_2(d - x + (\ell_2 + 1)e)$, or

(ii) $x \geq b$ and $c\mu_1 x \leq \frac{c^2(c+\ell_1\epsilon)}{(c+\epsilon)^2}\nu_1\nu_2(d - x + (\ell_2 + 1)e)$.

Arguing as before, if neither (i) nor (ii) holds, we decrease the value of x in the derivation $\mathcal{D}$ until one of them holds.

Suppose case (i) holds. This, followed by a use of (60), gives

$$\begin{aligned} \Big(\mu_1 + \frac{c(c + \ell_1\epsilon)}{(c+\epsilon)^2}\nu_1\nu_2\Big)b &\geq \frac{c(c + \ell_1\epsilon)}{(c+\epsilon)^2}\nu_1\nu_2(d + (\ell_2 + 1)e) \\ &\geq \frac{c(c + \ell_1\epsilon)}{(c+\epsilon)^2}\Big((c + \epsilon)\nu_1\mu_2 + \nu_1\nu_2\Big)b. \end{aligned} \tag{63}$$

Hence

$$\mu_1 \;\geq\; \frac{c(c + \ell_1\epsilon)}{c + \epsilon}\nu_1\mu_2.$$

We have $u' \geq c\mu_1 b$ by $x \geq b$ and (62). Combining this with (63) gives

$$u' \;\geq\; \frac{c}{c+\epsilon} \cdot \frac{c(c+\epsilon)(c+\ell_1\epsilon)\mu_1\nu_1\nu_2}{(c+\epsilon)^2\mu_1 + c(c+\ell_1\epsilon)\nu_1\nu_2}(d + (\ell_2+1)e).$$

The last two displayed inequalities suffice to prove that $\mathcal{D}$ is weakly subsumed by the triple $\langle \mu, \nu, \ell \rangle$ defined by (56). If B is (non-weakly) subsumed by $\langle \mu_2, \nu_2, \ell_2 \rangle$, then similar calculations lower bounding x' show that $\mathcal{D}$ is likewise (non-weakly) subsumed by the triple $\langle \mu, \nu, \ell \rangle$. We leave those details to the reader.

Finally, suppose case (ii) holds. Then we have

$$\Big(\mu_1 + \frac{c(c+\ell_1\epsilon)}{(c+\epsilon)^2}\nu_1\nu_2\Big)x \;\leq\; \frac{c(c+\ell_1\epsilon)}{(c+\epsilon)^2}\nu_1\nu_2(d + (\ell_2+1)e),$$

whence

$$d - x + (\ell_2+1)e \;\geq\; \frac{(c+\epsilon)^2\mu_1}{(c+\epsilon)^2\mu_1 + c(c+\ell_1\epsilon)\nu_1\nu_2}(d + (\ell_2+1)e). \qquad (64)$$

As before, $u' \geq c\mu_1 b$. In addition, using the inequality (62) with (61), we have

$$u' \;\geq\; c\Big(\frac{c(c+\ell_1\epsilon)}{c+\epsilon}\nu_1\mu_2\Big)b.$$

Further, again using (62), now with (64), yields

$$u' \;\geq\; \frac{c}{c+\epsilon} \cdot \frac{c(c+\epsilon)(c+\ell_1\epsilon)\mu_1\nu_1\nu_2}{(c+\epsilon)^2\mu_1 + c(c+\ell_1\epsilon)\nu_1\nu_2} \cdot (d + (\ell_2+1)e.$$

These three lower bounds on u' suffice to prove that $\mathcal{D}$ is weakly subsumed by the triple $\langle \mu, \nu, \ell \rangle$ as defined by (56). Similar calculations of lower bounds on x' show that if B is (non-weakly) subsumed by $\langle \mu_1, \nu_1, \ell_1 \rangle$, then $\mathcal{D}$ is (non-weakly) subsumed by the triple $\langle \mu, \nu, \ell \rangle$.

That completes the proof of Lemma 46. □

Lemma 46 lets us give a full characterization of when DTISP refutations exist, in terms of ABCD triples.

Theorem 47 *Any balanced, non-empty derivation $\mathcal{D}$ starting with a line with at least one quantifier is weakly subsumed by some* ABCD *triple.*

The proof of Theorem 47 is entirely analogous to the proof of Theorem 18; we leave the details to the reader. Likewise, the next theorems are proved entirely analogously to Lemmas 20 and 21 and Corollary 22. (Note, however, that we are not able to prove any useful analogue of Lemma 23 for DTISP refutations.)

Theorem 48 *Fix c and ϵ. There is a refutation if and only if there is a* ABCD*-triple $\langle \mu, \nu, \ell \rangle$ with $c\mu\epsilon < 1$ and $\rho(\mu, \nu, \ell) < 1$.*

Theorem 49 *Fix c and ϵ. There is a refutation if and only if there is a* ABE*-triple $\langle \mu, \nu, \ell \rangle$ with $c\mu\epsilon < 1$ and $\rho(\mu, \nu, \ell) < 1$.*

It is left to the reader to verify the details of the proofs.

References

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

[2] 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.

[3] L. Fortnow, R. Lipton, D. van Melkebeek, and A. Viglas, *Time-space lower bounds for satisfiability*, J. Association for Computing Machinery, 52 (2005), pp. 835–865.

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

[5] R. Kannan, *Towards separating nondeterminism from determinism*, Mathematical Systems Theory, 17 (1984), pp. 29–45.

[6] 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.

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

[8] I. Tourlakis, *Time-space tradeoffs for SAT and related problems*, Journal of Computer and System Sciences, 63 (2001), pp. 268–287.

[9] D. van Melkebeek, *Time-Space Lower Bounds for NP-Complete Problems*, World Scientific, 2004, pp. 265–291.

[10] ——, *A survey of lower bounds for satisfiability and related problems*, Foundations and Trends in Theoretical Computer Science, 2 (2007), pp. 197–303.

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

[12] R. Williams, *Inductive time-space lower bounds for SAT and related problems*, Computational Complexity, 15 (2006), pp. 433–470.

[13] ——, *Time-space tradeoffs for counting NP solutions modulo integers*, Computational Complexity, 17 (2008), pp. 179–219.

[14] ——, *Alternation-trading proofs, linear programming, and lower bounds*. Typeset manuscript. An 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, 2009.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de