CS 2429 - Propositional Proof Complexity

Lecture #12: 5 December 2002

Lecturer: Toniann Pitassi

Scribe Notes by: Toniann Pitassi

1 Introduction

In the last lecture, we completed the proof that any bounded-depth Frege proof of the propositional pigeonhole principle requires exponential size. In this final lecture, we will first discuss the best known upper bounds for the pigeonhole principle.

Then we will step back and review the relative complexities of the various proof systems discussed in this course: Resolution, Cutting Planes, bounded-depth Frege, Frege, LK, Extended Frege, and the Hajos Calculus. The next step is to prove superpolynomial lower bounds for Frege systems. This would be a major breakthrough in logic and complexity theory and seems to be beyond present reach.

We will discuss potential hard tautologies for proving superpolynomial lower bounds for Frege systems.

We will conclude with a discussion of other important open problems.

2 Upper Bounds for the Pigeonhole Principle

As we mentioned in an earlier lecture, Buss has shown the following theorem.

Theorem 1 (Buss). *For all n, there are polynomial-size (depth $O(\log n)$) Frege proofs of PHP_n^{n+1}.*

The proof of the above theorem is not difficult conceptually, but the proof is nontrivial. The idea is to define a multi-output polynomial-size formula that takes as input a binary vector, and outputs the number of 1's in the vector. This is well-known to be doable with carry-save addition circuits. Then using standard properties of addition, that can be efficiently proven about the carry-save addition circuits, it is shown that from the pigeon assumptions of $\neg PHP_n^{n+1}$, that the sum of all of the variables $P_{i,j}$ is at least $n+1$, and from the hole assumptions, that the sum of this same set of variables is at most n, and hence we reach a contradiction.

For the weak pigeonhole principle, Paris Woods and Wilkie showed that there are quasipolynomial-size bounded-depth proofs. Recently their result was improved by Maciel, Pitassi and Woods who give a proof of depth that is only .5. This means that each formula in the LK derivation is either a conjunction of *polylogn* many literals, or a disjunction of *polylogn* many literals.

Theorem 2. *For all n, $PHP_n^{n^2}$ has depth .5 LK proofs of size $n^{O(\log n)}$, and PHP_n^{2n} has depth .5 LK proofs of size $n^{O(\log n)^2}$.*

3 Relative Complexity of Proof Systems

Recall the proof system hierarchy from Lecture 2. In previous lectures, we have proven exponential lower bounds for Resolution, Cutting Planes and bounded-depth Frege proofs.

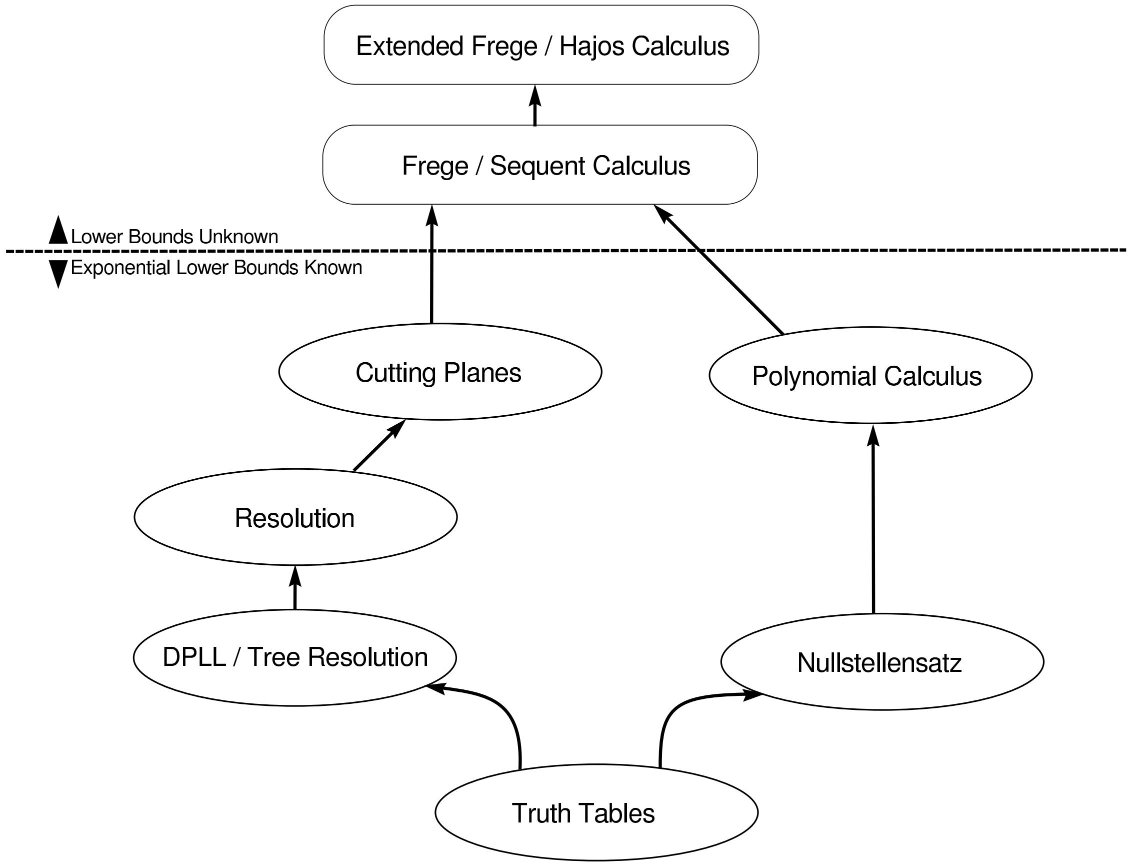

Figure 1: Proof System Hierarchy

The most outstanding open problem is to extend the lower bounds to Frege systems. Currently, the best size lower bound for Frege systems is the following theorem, due to Sam Buss.

Theorem 3 (Buss). *There is a family of tautologies that require an $O(n)$ line Frege or Extended Frege proof, and an $O(n^2)$ size Frege or Extended Frege proof.*

The family of tautologies are very simple. They are of the form $F \vee (F \vee (F \vee ...(F \vee T)...)$. The idea behind the lower bound is simple. The formula is a tautology but it is not an instance of a shorter tautology. Therefore, each connective must be "peeled" off by some rule. Since each rule involves only a finite number of active connectives, the total number of rules must be linear in the size of the tautology. The size bound is proven by formalizing the fact that the rule must break apart the statement from outside to inside and indoing so, it ends up repeating subformulas many times.

Although substitution Frege (sF) is polynomially equivalent to Extended Frege, it is substantially harder to prove lower bounds on the length of sF proofs. The best lower bound for sF is due to Urquhart, who proved that there is a family of formulas that require an $O(n/\log n)$ line sF proof. The proof does not give an explicit family of formulas. Proving quadratic lower bounds on the symbol size of an sF proof is also an open problem.

4 Potential Hard Examples for Frege

4.1 Partial Consistency statements

The classic example of statements that are unprovable in systems of arithmetic are consistency statements. Ordinary consistency statements are not expressible propositionally and hence they do not apply to our setting. However, it is possible to come up with a partial consistency statement that *is* formalizable propositionally. The idea here is that there is a parameter n, and $Con_S(n)$ expresses that there is no S-proof of False with size at most n.

Cook was the first to formalize partial self-consistency statements for EF, $Con_{EF}(n)$. He showed that there are polynomial-size EF proofs of $Con_{EF}(n)$. Later, Buss studied partial consistency statements for Frege as well, and he proved that there are short Frege proofs of $Con_{Frege}(n)$.

Below we will outline Buss's formalization of $Con_S(n)$ for a propositional Frege system S. In order to formalize $Con_S(n)$, we have to show how to represent an S-proof as a bit sequence, and show how to define formulas for interpreting the sequence as a sequence of formulas.

Proofs in S will be represented by words/strings in a 19 character alphabet Σ containing p, 0, 1, parenthesis, comma and 13 propositional connectives. A propositional variable p_i will be represented by "p" followed by a string of 0's and 1's coding i in binary. Commas will be used to separate formulas in a proof. Strings over Σ are further encoded over the language $\mathcal{L} = \{0, 1\}$ by assigning a unique 5-bit code to each symbol in Σ. Thus an S-formula with k symbols (counting the symbols used to code the subscripts of the variables) will be coded by $5k$ truth values.

Let $\mathbf{x}$ represent a vector of propositional variables $x_1, \ldots, x_{5k}$. We want to show that "$\mathbf{x}$ codes a formula" can be represented by a polynomial-size propositional formula. The key tool to parsing a formula coded by $\mathbf{x}$ is to show how to count parenthesis. Let $|\mathbf{x}|$ denote the number of logical symbols in $\mathbf{x}$. Then "$\mathbf{x}$ codes a formula" will be expressed as: (1) ...

Now we need to define the concept of an S-proof with a polynomial-size formula. A proof is coded by formulas separated by commas, where each formula is either an instance of an axiom scheme or follows by modus ponens from two previous formulas. For S there are a finite number of axiom schemes so it is easy to define a polynomial-size formula expressing "$\mathbf{x}$ is an instance of an axiom scheme." Similarly it is possible to define a polynomial-size formula expressing "$\mathbf{x}$ is derived from $\mathbf{y}$ and $\mathbf{z}$ by modus ponens."

Buss proved, using clever formulas for evaluating the truth value of a formula on a given input, that $Con_{Frege}(n)$ is provable with polynomial-size Frege proofs. He also showed that $Con_{EF}(n)$ is provable with polynomial-size Frege proofs if and only if Frege can polynomially simulate Extended Frege. Thus $Con_{EF}(n)$ is a potential hard example for Frege.

Partial consistency formulas are excellent from one point of view: if there are hard examples at all separating EF from Frege, then we know that Con_{eF} is such an example. However, a disadvantage of these formulas is that there is no to believe that they are hard in the first place, other than the belief that EF is stronger than Frege. So these examples seem to provide no further evidence that Frege is not polynomially bounded.

In an attempt to rectify the above criticism, Avigad reformulates partial consistency statements for Frege and Extended Frege systems in an extremely elegant way, by using a particularly simple formulation of a Frege system over the single NAND connective. He defines a Frege system $\mathcal{F}_1$ over this one connective consisting of only one axiom:

$$nand(\overline{g}, \overline{f}, nand(\overline{f})),$$

and one rule:

$$\frac{nand(\overline{g}, \overline{f}),\ nand(\overline{g}, nand(\overline{f}))}{nand(\overline{g})}$$

and proves that $\mathcal{F}_\infty$ is complete.

Because of the simplicity of $\mathcal{F}_\infty$, he can relate the consistency of $\mathcal{F}_\infty$ to a combinatorial property of hereditarily finite sets.

4.2 Matrix identities

There are several propositional statements that come from linear algebra that do not seem to have feasible Frege proofs. The simplest of these (due to Steve Cook) expresses the fact that for any two n by n matrices A and B,

$$AB = I \rightarrow BA = I.$$

To express this propositionally, we consider 0/1 matrices A and B over GF_2. For each n, there are $2n^2$ underlying variables $a_{i,j}$ and $b_{i,j}$ describing the matrices. It is not hard to express this identity in these variables with polynomial-size formulas, or by introducing a linear number of auxillary variables, one can obtain a CNF formula of polynomial-size. for each n. We will denote these tautologies by AB_n.

What is known about the complexity of these tautologies? Bonet, Buss and Pitassi originally conjectured that these tautologies should have EF proofs of polynomial-size, and Frege proofs of quasipolynomial-size. The reasoning behind this conjecture is based on the fact that iall concepts in the standard proof are expressible as quasipolynomial-size formulas. Therefore it is reasonable to expect that the relevant properties of these concepts (determinant, etc) can also be proven in quasipolynomial-size, to give a Frege proof of quasipolynomial-size, simulating the standard proof.

However, despite considerable effort, there is still no known subexponential-size Frege proof of these tautologies to date! The best known proof, due to Cook and Soltys, is a polynomial-size Extended Frege proof. In a systematic study of matrix identities (including this one), Cook and Soltys developed a quantifier-free, three-sorted logical theory for linear algebra called LA. The sorts are indices, field elements and matrices. In LA one can express universal matrix identities such as the one above, and prove all of the ring properties of matrices. Further, formulas in LA translate into familes of propositional formulas. Using the language of LA, they formalize Gaussian elimination enabling them to give polynomial-size EF proofs of AB_n, along with short proofs for many other matrix identities. They also put the identities into equivalence classes (based on how much proof-theoretic strength beyond LA seems to be required to prove them) and give LA proofs of equivalences within each class.

4.3 Circuit Lower Bounds

In this section we will explain one way of expressing that SAT (or some other NP-hard language) does not have polynomial-size circuits with a family of propositional formulas.

Let C_n be an n input circuit of size s. We'll typically assume that s is superpolynomial, and here we'll think of $s = n^{\log n}$. $C_{n,s}$ can be represented by $s^{O(1)}$ variables. (In this case $n^{O(\log n)}$ boolean variables), $c_1, \ldots, c_q$. Let f be an NP-complete language (such as SAT), and let f_n be f on inputs of length n. We will think of f as being fixed, and each f_n will be viewed as a particular, fixed, binary string of length 2^n, $f_n = \alpha_1\alpha_2\ldots\alpha_{2^n}$. Then we can express that f_n cannot be computed by $C_{n,s}$ by the propositional statement, $Hard(f)_{n,s}$, as follows.

$$\bigvee_{i,\alpha_i=1} C_{n,s}(i) = 0 \vee \bigvee_{j,\alpha_j=1} C_{n,s}(j) = 1$$

where $C_{n,s}(i) = 0$ is a propositional formula (in $c_1 \ldots c_q$) that is true if and only if the circuit $C_{n,s}$ coded by $c_1, \ldots, c_q$ outputs 0 on input i.

In order to formulate $Hard(f)_{n,s}$ as a CNF formula, it is necessary introduce a linear number of extra variables in addition to $c_1, \ldots c_q$ in order to simplify the description of the formulas $C_{n,s}(i) = 0$, and $C_{n,s}(i) = 1$ so that they can be expressed in CNF form. This can be done in many standard ways.

The formula $Hard(f)_{n,s}$ has length $2^{O(n)}$, and $s^{O(1)}$ many variables (Here, $2^{O(\log n)}$ many variables.) If in fact $HardD(f)_{n,s}$ is true, then a straightforward tree-like Resolution refutation of $\neg Hard(f)_{n,s}$ will be exponential in the number of variables, and hence is not polynomial in the size of $\neg Hard(f)_{n,s}$.

It has been conjectured that $Hard(f)_{n,s}$ requires superpolynomial-length Frege proofs for f an NP-complete language, and $s = n^{\log n}$. Of course, if $Hard(f)_{n,s}$ is not true, then the conjecture is vacuous. But if f really does require large circuits, then lower bounds for $Hard(f)_{n,s}$ in certain proof systems are important since they shed light on the metamathematical properties of proving circuit lower bounds. Recently it has been shown by Razborov that $Hard(f)_{n,s}$ requires superpolynomial-size Resolution proofs for many values of s.

Lower bounds for $Hard(f)_{n,s}$ are open not only for Frege, but also for Cutting Planes and bounded-depth Frege systems. Note that this is a family of formulas that is parameterized not only by n (the size of the input to f) but also by s, the size of the circuit, $C_{n,s}$. When s is small enough (say sublinear), then $Hard(f)_{n,s}$ is known to be true, and when s is extremely large (say 2^n) then $Hard(f)_{n,s}$ is known to be false since there are exponential-size circuits for computing any boolean function. Thus as s decreases, the formula $Hard(f)_{n,s}$ becomes "more true" and therefore should be easier to prove. Notice also that when s decreases then the number of variables underlying $Hard(f)_{n,s}$ gets smaller, so the obvious tree-like Resolution proof does get smaller.

Random (unsatisfiable) kCNF formulas are another example of a family of formulas that are parameterized by two values, n, the number of variables, and m, the number of clauses. When n is fixed, as m increases, a random formula with n variables and m clauses is "more false" and hence should be easier to refute.

Both of these families of formulas (random formulas and formulas expressing the hardness of computing a specific NP-hard function) are generally believed to be hard for Frege systems for *some* nontrivial value of the parameters. However, there is no real evidence that I know of indicating that this should be the case.

4.4 Number theoretic statements

Another rich source of hard examples comes from number theory.

Let $Composite(a)$ be a propositional formula stating that there exist two numbers $1 < u, v < a$ such that $uv = a$. Then for each prime number p, let $Prime_p$ be the propositional formula $\neg Composite(p)$. Jan Krajicek has suggested that this family of tautologies should be hard for certain primes. Charlie Rackoff conjectures that it is hard for *every* prime.

5 Open Problems

- Is there a strongest proof system? Give some evidence one way or the other.
- Prove that random kCNF formulas are hard for AC_0-Frege.
- Prove that the Tseitin graph tautologies are hard for Cutting Planes.
- There are several proof systems along the lines of Cutting Planes that are important and relatively unexplored. One example is the Lovasz-Schriver Proof System, which is based on 0/1 programming. The initial inequalities are like those for Cutting Planes. However, now one can substitute x for x^2 anywhere. Also the division rule is not present. Non-negative degree 2 polynomials can be obtained by multiplying two non-negative linear quantities or by adding the square of any linear quantity. This system polynomially simulates resolution and can also prove the propositional pigeonhole principle efficiently. It has feasible interpolatin, and hence it is known to be not polynomially bounded under the assumption that NP is not contained in $P/poly$. However there is no explicit hard tautuology known for it. Another question is whether or not this system can efficiently simulate Cutting Planes.
- Does the matrix identity $AB = I \rightarrow BA = I$ have quasipolynomial-size Frege proofs?
- How do the various plausibly hard tautologies that we discussed above compare with one another. For example, how does AB_n compare with $Con_{EF}(n)$? Can Frege plus AB_n efficiently prove $Con_{EF}(n)$? Can AC_0 Frege plus AB_n prove $Con_F(n)$? Give a systematic treatment of the relative complexities of these various classes of examples.

6 Bibliographic notes

- Sam Buss. Polynomial-size proofs of the propositional pigeonhole principle." Journal of Symbolic Logic, pages 916-927, 1987.
- J.B. Paris, A.J. Wilkie, and A.R. Woods. "Provability of the pigeonhole principle and the existence of infinitely many primes." Journal of Symbolic Logic. pages 1235-1244, 1988.
- Alexis Maciel, Toniann Pitassi, and Alan Woods. "A New Proof of the Weak Pigeonhole Principle."
- Sam Buss. "Some remarks on the lengths of propositional proofs." Archive for Mathematical Logic, Volume 34, pages 377-394, 1995.

- Stephen A. Cook. "Feasibly Constructive Proofs and the Propositional Calculus," In Proceedings of 7th Annual Symposium on Theory of Computing, pages 83-97, 1975.
- Sam Buss. "Propositional Consistency Proofs."
- Jeremy Avigad. "Plausibly Hard Combinatorial Tautologies."
- Stephen A. Cook and Michael Soltys. "The Proof Complexity of Linear Algebra." Proceedings LICS 2002.
- Michael Soltys. "The complexity of derivations of matrix identities." PhD Thesis, University of Toronto, 2001.