The Complexity of Some Geometric Proof Systems

ABDUL AZIZ SAUD ABDUL GHANI

DURHAM UNIVERSITY

The Complexity of Some Geometric Proof Systems

Abdul Ghani

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

Algorithms and Complexity in Durham
Department of Computer Science

February 2023

Contents

Declaration of Authorship

The work in this thesis is based on research carried out in the Algorithms and Complexity Group (ACiD) at the Department of Computer Science, Durham University, England. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

I'm grateful for my friends. I want to thank Luke, for a lot, James and Ace, for all the music and the farmers walks, Tom, for being so incredible, and for being so insightful, and Ecem, for pointing out (correctly) that there's something cool about the phrase '$\lambda \in \Lambda$', and also for introducing me to her brother.

Out of my research group I want to thank Dave (or David) for being my Dungeon Master, and for telling me (also correctly) that It's All Going To Be Okay, and Karl, Siani, for tolerating certain things I had to say.

I'm also grateful for my family. In particular I'm grateful for my two brothers.

I would like to thank Nicola Galesi very much for his fruitful collaboration in the later years of my PhD. And of course I am most indebted to my supervisors, without whom I wouldn't be here. I'm thankful for Stefan introducing me to proof complexity as an undergraduate, and for Barny, for all his hard work in supervision and collaboration, but also for his sense of humour, and most importantly, for his apparantly unshakeable belief that I would finish this thesis at all.

For Brishl

Chapter 1

Introduction

The research contained here is housed within the field called proof complexity.
Here we are concerned with the complexity of refuting, i.e. proving the unsatisfiability of, statements given in some formal language. More specifically, we are in the realm of propositional proof complexity, and we try to understand the complexity of proving true statements given in propositional (quantifier free) language.

There are many different measures of the 'complexity of a proof'. We have to first define what it means to be a proof at all, and this is done by fixing some given axioms and some permitted rules of inference. Then we must pick a measure of complexity – we could, for example, investigate the number of lines in a proof, or maybe, the total number of characters needed to specify the proof, perhaps, the number of applications of a certain inference rule used in any given proof, usually the inference rule that is most computationally costly to simulate - there are many such metrics that could be said to capture a sort of complexity.

We have much choice and making different choices gives us differing readouts of complexity. Propositional proof complexity theorists enlighten us to the consequences of these decisions by relating these differing ideas of complexity. And by linking these ideas together, they create an impression of what the 'true, inherent complexity' of proving a certain thing should be.

The general definition of a proof system given in the seminal paper by Cook and Reckhow [20] is a function from strings in some alphabet *onto* the set of all propositional tautologies that is computable in time polynomially bounded in the size of its input. The intended meaning is that f takes as an input a proposed proof and produces as an output the tautology demonstrated by the proof. Then the function f being surjective means that it is complete, as any tautology is reached by at least one proof, and its

being computable in polynomial time means that a proof can be vetted efficiently.

We can now define precisely our first notion of the complexity of a proof of a tautology T relative to some proof system f as the bitsize of the smallest proof:

$$\min |p| : f(p) = T.$$

The proof system f is said to be *polynomially bounded* if there exists some polynomial p such that for any tautology T there exists an x such that $T = f(x)$ satisfying $|x| \leq p(|T|)$. It is quickly seen that a polynomially bounded proof system exists if and only if **NP** = **co-NP** - as the problem of deciding tautologicity is **co-NP** complete, and the (apparently always polynomially sized) proof of tautologicity can be nondeterministically divined by a nondeterministic Turing machine - therefore proving that no such system can exist will separate **NP** from **co-NP** and and so **P** from **NP**. This is one of the main original motivations behind the field of propositional proof complexity.

A proof system f *p-simulates* another proof system g if there exists a polynomial time computable function t such that $g(x) = f(t(x))$ for all strings x. We can now prove that two proof systems are polynomially equivalent, in the sense that they polynomially simulate each other, or we could prove that they are incomparable, in the sense of no such simulation existing in either direction, or we could exponentially separate the two, by showing that one system has only exponentially sized proofs for some family of tautologies exhibiting efficient (polynomially sized) proofs in the other. A major open question concerns the existence of a *p-optimal* proof system - that is, a proof system that p-simulates every other proof system. The resolution of this question has a number of complexity and proof-theoretic implications. For example, showing the existence of such a system would imply the existence of a language complete for the intersection **NP** $\cap$ **co-NP**, whereas proving that such systems dont exist would separate **NEXP** from **co-NEXP** [47, 56].

Classifying and comparing these algorithms comes down to proving separation and proving p-simulation. Proving efficient simulation is (usually) relatively straightforward and the majority of effort in proof complexity is spent on demonstrating separation. This is most often done by showing upper and lower bounds for 'benchmark' tautologies, which we turn to next.

1.1 Combinatorial contradictions

Here we define and collect a few of the tautologies appearing most often in proof complexity. Their relevance will be revealed in the next section, where they will be used to demonstrate the relative strengths of the proof systems appearing in this thesis.

It might be startling to see the following contradictions described as tautologies. But note that to prove that some ϕ is a tautology is the same as to prove its negation $\neg\phi$ is a contradiction. Therefore (despite their being opposites) we will use 'tautology' and 'contradiction' interchangeably throughout this thesis, which is completely customary in the field of proof complexity.

1.1.1 Uniform generation of propositional contradictions

Definition 1.1. A *language* is a set of non-logical atomic symbols with an associated arity. Zero arity is allowed, and in this case the symbol is simply a constant. Arity greater than zero corresponds to a relation.

Definition 1.2. A *first order (FO) formula* over a language $\mathcal{L}$ is any of the following:

1. An atomic symbol $R(\overline{x})$, where $R \in \mathcal{L}$, and $\overline{x}$ is a tuple of length the arity of R, consisting of free variables and constants from $\mathcal{L}$,
2. $\phi_1 \square \phi_2$, where $\square$ is any boolean logical connective from $\{\wedge, \vee, \implies\}$, and ϕ_1, ϕ_2 are FO formulae,
3. $\neg\phi$, ϕ a FO formula, or
4. $Qx\phi$, where $Q \in \{\forall, \exists\}$ is a quantifier and ϕ is a FO formula. In this case, we say every appearance of x in ϕ is *bound*.

We say that a FO formula is a *sentence* if every variable appearing in that formula is bound.

Definition 1.3. A structure $\mathfrak{M}$ of some language $\mathcal{L}$ is a pair, consisting of (1) a *domain of discourse* D being any set containing the constant symbols from $\mathcal{L}$, and (2) an *interpretation* of every nonlogical symbol in $\mathcal{L}$, being an assignment of true or false to every instantiation of every relation $R \in \mathcal{L}$ with elements from D of the appropriate arity.

Such a structure is said to be a *model* of some FO formula $\mathcal{T}$ in the language of $\mathcal{L}$ if it satisfies that formula, where any unbound variables are taken to be universally quantified.

In this section we show how to uniformly (that is, computably) generate from any FO sentence $\mathcal{T}$, a propositional (quantifier-free) claim that it has a model of some finite size, in the manner first introduced by [78].

We are given some FO sentence $\mathcal{T}$ consisting of (without loss of generality prenex) formulas

$$\forall x_1 \exists y_1, \ldots, \forall x_k \exists y_k \; F(x_1, \ldots, x_k, y_1, \ldots, y_k), \tag{1.1}$$

where F is quantifier free and in CNF. For any natural number $n \in \mathbb{N}$ we can uniformly produce a system of propositional formulae which are feasible if and only if Equation (1.1) has a model of size n. This transformation proceeds along the following steps:

Firstly we eliminate the existential clauses with a sort of Skolemization. Introduce for each existentially quantified y_i a *Skolem variable* $S_i(x_1, \ldots, x_i, y_i)$. Ask that a witness to the existential demand always exists in the finite domain of discourse $\{1, 2, \ldots, n\}$ by including the *Skolem clauses*

$$\bigwedge_{x_1, x_2, \ldots, x_i = 1}^{n} \bigvee_{y_i = 1}^{n} S_i(x_1, \ldots, x_i, y_i)$$

asking that every $x_1, \ldots, x_i$ taking values from $[n]$ has a witness in $[n]$ (here $[n] = \{1, 2, \ldots, n\}$). Express Equation (1.1) in the now purely universal form

$$\forall x_1, \ldots, x_k, y_1, \ldots, y_k \left(\bigwedge_{i=1}^{k} S_i(x_1, \ldots, x_i, y_i) \right) \implies F(x_1, \ldots, x_k, y_1, \ldots, y_k). \tag{1.2}$$

We then eliminate the remaining universal quantifiers by replacing Equation (1.2) with the n^{2k} clauses resulting from instantiating the x, y with all possible values from $[n]$. This process concludes with a propositional statement in CNF, and it is plain to see that the output is satisfiable if and only if the input had a finite model of size n.

We now exemplify this transformation by deriving some the most famous principles appearing in propositional proof complexity, which will appear many times throughout this thesis.

1.1.2 Ordering Principles

The *Ordering Principle* (OP) asks for an order with no minimal element. It is the conjunction of the following sentences, over a language containing a binary relation '$<$',

written here as infix:

$$\begin{gathered}\textit{self:}\ \forall x\ \neg(x < x)\\ \textit{trans:}\ \forall x, y, z\ (x < y) \wedge (y < z) \implies (x < z)\\ \textit{lower:}\ \forall x \exists y\ (y < x).\end{gathered}$$

The *Linear Ordering Principle* (LOP) asks further that this order is total, or linear, by also including the following axiom:

$$\textit{total:}\ \forall x, y\ (x < y) \vee (y < x).$$

This is, of course, not yet a contradiction - the integers are an example model. However there can be no finite model, and so the translation of these sentences produces the following contradiction, P_{xy} having the intended interpretation $x < y$, and the $S_{i,j}$ are the Skolem variables resultant from the transformation just described:

$$\begin{gathered}\neg P_{i,i} \qquad \forall i \in [n]\\ \neg P_{i,j} \vee \neg P_{j,k} \vee P_{i,k} \qquad \forall i, j, k \in [n]\\ \neg S_{i,j} \vee P_{i,j} \qquad \forall i, j \in [n]\\ \bigvee_{i \in [n]} S_{i,j} \qquad \forall j \in [n].\end{gathered}$$

1.1.3 The Pigeonhole Principle

The *Pigeonhole Principle* PHP is the conjunction of the following FO sentences, over the single binary relation P and constant 1:

$$\begin{gathered}\textit{holed:}\ \forall x\ \exists y P(x, y)\\ \textit{inj:}\ \forall x, y, z,\ x = y \vee \neg P(x, z) \vee \neg P(y, z)\\ \textit{first:}\ \forall x \neg P(x, 1)\end{gathered}$$

The translation gives us something basically equivalent to the following propositional axioms, which we call the PHP_n:

$$\begin{gathered}\textit{holed:}\ \forall x \in [n] \bigvee_{y \in [n]} P_{xy}\\ \textit{inj:}\ \forall x \neq y \in [n], z \in [n] \setminus \{1\}\ \neg P_{xz} \vee \neg P_{yz}\\ \textit{first:}\ \forall x \in [n]\ \neg P_{x1}\end{gathered}$$

We say basically equivalent because here, and elsewhere, we have made a simplification in the translation just given in Section 1.1.1: when taken literally, we should have introduced a Skolem variable S_{ij}, and the jth existential *holed* axiom should actually become $\bigvee_{i=1}^{n} S_{ji}$. However, Equation (1.2) would also give us $S_{ji} \implies P_{ji}$, and as propagating this Horn clause is done at no real complexity cost in all the proof systems and metrics we consider in this thesis, it makes no difference to apply it a priori to the principle at hand.

1.1.4 Tseitin contradictions

The Tseitin contradictions are not necessarily generated from FO sentences, but nevertheless, are just as important. They are important historically (see, eg, Section 1.2.1 in the sequel), but perhaps more usefully they are also important practically, as the structure of each contradiction as a propositional formula reflects the structure of the graph from which it is generated. For example, we will see in Chapter 2 that a parameter of the graph, the so called *circuit rank*, tells us something about the complexity of the Tseitin formula that we generate from it.

Definition 1.4. Let $G = (V, E)$ be a graph and ω a *charging function* $\omega : V \to \{0, 1\}$ that is *odd*, that is, $\sum_{v \in V} \omega(v) = 1 \mod 2$. The *Tseitin contradiction* $\mathsf{Ts}(G, \omega)$ is the CNF resultant from the translation of the parity constraint

$$\sum_{\substack{e \in E \\ e \ni v}} x_e = \omega(v) \mod 2. \tag{1.3}$$

We have one such constraint for every $v \in V$, and the variables x_e range over the edges $e \in E$.

That this is indeed a contradiction is just the handshaking lemma - if we sum up all the neighbourhoods of all nodes, each edge is counted exactly twice. Here we collect some facts that will be used later in Chapter 2. The following trick is due to [85] but we include the proof as it is instructive.

Lemma 1.1. *Let b be a boolean assignment to the variables in $\mathsf{Ts}(G, \omega)$, where G is some connected graph. If b falsifies the parity constraints for exactly some set of vertices $V' \subseteq V(G)$ with $|V'| \geq 2$, then for any $v_1, v_2 \in V'$, we can find a boolean assignment b' falsifying the parity constraints only for the set of vertices $V' \setminus \{v_1, v_2\}$.*

Proof. We simply take any path p connecting two distinct vertices v_1 and v_2 in G and flip the assignment b gives to the edges in p. The endpoints see their parity flipped, as they

are adjacent to exactly one flipped edge, and their previously violated parity constraints are now satisfied. The vertices en route on p between v_1 and v_2 have unaffected parity, as they are adjacent to exactly two flipped edges. And the remaining vertices not on p altogether have their incident variables completely untouched. □

Corollary 1.1. *Let ω be an odd charging of some graph G. For any $v \in V(G)$, we can find a boolean assignment b falsifying only the parity constraint for v.*

Proof. Pick any boolean assignment that falsifies the parity constraint at v. There must be an odd number of nodes B with violated parity constraint, as otherwise (by the previous fact) we can find a boolean assignment satisfying all the constraints of what was meant to be a contradiction.
So, while $|B| \geq 3$, pick two $v_1, v_2 \neq v \in B$ and apply the previous fact. □

1.2 Proof systems

In this section we describe the proof systems that are addressed in this thesis. There are, of course, excellent surveys on the topic of these systems and on proof complexity in general - see, for example, [57, 58, 80] for more on Sums Of Squares, [36] for a more recent survey focussing on Tseitin contradictions and the more recent Stabbing Planes proof system, and [8] for a general view of proof complexity as a whole.

1.2.1 Resolution

Resolution is without a doubt the oldest and most well studied propositional proof system. It is also the only proof system discussed in this thesis that is not inherently 'geometric' (in a sense defined after this subsection), but rather it works directly against the CNF in question.
It was introduced by Blake in 1937 in his PhD thesis ([11]) and modernised as a refutation system in a series of papers by Davis, Putnam, Logemann, and Loveland [30, 31]. The DPLL algorithm, which to this day forms the basis of many widely used SAT solvers, is essentially an implementation of Resolution.

Definition 1.5. The *Resolution* proof system has the single inference rule (the 'Resolution rule')

$$A \vee x,\ B \vee \neg x \implies A \vee B$$

where A, B are clauses and x a variable. A sequence of clauses $R_1, \ldots, R_k$ is a Resolution refutation of a set of clauses (axioms) $C_1, \ldots, C_a$ if $R_k = \bot$ is an empty clause and each

R_i is either some axiom C_j or is derived from two previous R_l, R_m $(l, m < k)$ by the Resolution inference rule. The *size* of the Resolution refutation is k.

The refutation is said to be *treelike* if each R_i that is derived by an application of the Resolution rule itself only appears once as an antecedent in an application of the Resolution rule.

The meaning of treelike comes from viewing the refutation, as we will often do, as a graph, with a directed edge from the two antecedent clauses to the consequent clause. Then this graph is acyclic, with every non-axiom clause having indegree exactly two, but technically unlimited outdegree (in the non-treelike case).

Resolution is clearly sound - as the Resolution rule preserves satisfiability and unsatisfiability - and its completeness can be easily shown by induction on the number of variables. The proof is roughly as follows: given a set of contradictory clauses C, pick any variable x appearing in C and partition C into three sets P, N, A, where x appears positively, negatively, and is altogether absent, respectively. If one of P or N is empty, x is a pure literal and can be removed from C without affecting its satisfiability. Otherwise, we resolve all pairs of clauses from P and N finding a contradiction missing x entirely [45].

The first lower bounds for Resolution were given by Tseitin in 1968 [84], for a principle that became eponymous - the Tseitin principle defined just above in Section 1.1.4. He gave there a subexponential size lower bound. This lower bound for Tseitin was later improved to an exponential lower bound for regular Resolution, which is a weakened form of Resolution where each variable is resolved upon at most once on any path from a clause to the contradictory root of the refutation, by Galil in his PhD thesis [38]. However, the first exponential lower bound for size in general Resolution, and arguably the most famous result in Propositional Proof Complexity as a whole, was given by Haken [45] in 1985, for a different principle:

Theorem 1.1 ([45], Section 2)**.** *The* PHP_n *requires size* $2^{\Omega(n)}$ *to refute in Resolution.*

A crucial innovation in that paper, reflected in many subsequent Resolution lower bounds (including our own in Chapter 4) is that if one aims to show a size lower bound, one might benefit from first demonstrating a *width* lower bound, and then showing that a width lower bound implies a size lower bound.

Definition 1.6. The width of some clause $x_1 \vee x_2 \vee \ldots \vee x_w$ made up of the literals x_i is w, and the width of a refutation $R_1, \ldots, R_k$ is the maximum width of the clauses R_i.

Typically a width lower bound is easier to show than a direct size lower bound and is often given by a relatively straightforward prover-adversary argument (see e.g. [72]). Sometimes this is even enough, for Resolution enjoys a size-width tradeoff:

Theorem 1.2 ([9], Theorem 3.5)**.** *Let $C := C_1, \ldots, C_a$ be a set of contradictory axioms over n variables and maximum width w_a. Let w be the minimum width of a Resolution refutation of C and s the minimum size. Then*

$$w \leq w_a + \sqrt{n \log(s)}.$$

It turns out that this tradeoff is tight. In [14], the following is proven:

Theorem 1.3 ([14], Theorem 3.1)**.** LOP_n*, even when converted into 3-CNF, requires width $\Omega(n)$ to refute in Resolution, but still has refutations of polynomial size.*

1.2.2 Geometric proof systems

Definition 1.7. A set $X \subseteq \mathbb{R}^n$ is called *semialgebraic* if it is exactly the set of solutions to some finite set of polynomial inequalities $q_1 \geq 0, q_2 \geq 0, \ldots, q_k \geq 0$. If the q_i are linear, so X is an intersection of halfspaces, we call X a *polytope*.

A crucial theme appearing often in proof complexity is that, given some CNF sentence C over propositional variables V, we can find a set of linear inequalities P over continuous variables V', whose simultaneous binary solutions (that is, every variable substituted with a 0 or 1) correspond exactly to satisfying assignments of C. The transformation itself is straightforward: for every clause $x_1 \vee x_2 \vee \ldots \vee x_p \vee \neg y_1 \vee \ldots \vee \neg y_n$ over variables $x_i, y_j \in V$ we emit the inequality $\sum_{i=1}^{p} x'_i + \sum_{j=1}^{n}(1 - y'_j) \geq 1$, over the primed variables $x'_i, y'_j \in V'$. We also emit the bounds $0 \leq v' \leq 1$ for all $v' \in V'$. This means that any algorithm capable of proving infeasibility of an Integer Linear Program (ILP) is now a proof system - instead of proving unsatisfiability of the CNF formula C, we instead prove integer freeness of P, or equivalently, the emptiness of the integer hull $R \subseteq P$ (see e.g., [43]). In this context we will call these ILP solvers *geometric proof systems*, for that is what they now do.

These algorithms often work by constructing relaxations R_t of R parameterized by some tightness (or rank) parameter $t \in O(|V|)$, with the properties

1. (Monotonicity.) $R_{t+1} \subseteq R_t$.
2. (Soundness.) If R_t is empty for some t, then R is empty.

3. (Completeness.) If R is empty, then R_t is empty for some t.

These R_t will be linear or semidefinite programs, or something otherwise much easier to solve than the original ILP. If $\mathbf{P} \neq \mathbf{NP}$ then the time-complexity of producing the relaxation R_t is at least exponential in the parameter t, and one proves complexity lower bounds for the procedure by proving nonemptiness of R_t for the largest possible t.

1.2.3 Sums Of Squares over the boolean hypercube

The *Positivstellensatz* proof system is one such geometric proof system. If we are promised, for some finite set of polynomials p_i, q_i, that $p_1 \geq 0, p_2 \geq 0 \ldots p_m \geq 0, q_1 = 0, q_2 = 0 \ldots q_n = 0$ (over variables restricted to 0/1) we can infer that

$$\sum_{I \subseteq \{1 \ldots m\}} \left(\left(\prod_{i \in I} p_i \right) \left(\sum_j (a_{I,j})^2 \right) \right) + \sum_{i=1}^{n} b_i q_i \geq 0 \tag{1.4}$$

where the $a_{I,j}$ and b_i are arbitrary polynomials and the empty product is read as 1. This is because the nonnegativity of the p_i implies the nonnegativity of their products, which are then multiplied by the sums of squares $\sum_j (a_{I,j})^2$, which are all nonnegative everywhere as we are only working with real numbers, and finally, each q_i multiplied by anything at all gives 0, and therefore so does their sum. The given set of polynomial inequalities is refuted if we derive an obvious contradiction such as $-1 \geq 0$, in which case we call (1.4) a *Positivstellensatz refutation.* The *Sum Of Squares* (SOS) proof system is the special case where the indices I under the first summation are restricted to singletons.
The system is obviously sound. (That the system is essentially complete is not as obvious, and this will discussed below.) The *degree* of the refutation is the maximum degree of the polynomials $(a_i)^2, (b_{I,j})^2 \prod_{i \in I} p_i, c_i q_i$, and will be the complexity measure focused on later in this thesis.

Sum Of Squares based techniques have recently come to the interest of researchers due to its potential use in a refutation of the unique games conjecture [6]. As will be seen later, SOS based algorithms are optimal for a variety of problems assuming the UGC - so beating SOS would refute the UGC.

We might also consider a dynamic system with the following rules:

1. $f^2 \geq 0$ for any polynomial f.

2. $a = 0 \implies a \geq 0$.

3. $a \geq 0, b \geq 0 \implies va + wb \geq 0$ for any polynomials a, b and SOS v, w.

4. $a = 0, b = 0 \implies va + wb = 0$ for any polynomials a, b, v, w.

It is worth pointing out that the $\implies$ here is meant as a syntactic production rule and not logical implication, although, as the production rules are sound by inspection, the logical implication is valid (over the reals).

We can convert a dynamic proof into a static one by induction on the height of the tree. In the case of production rule 3, we have two SOS derived polynomials

$$a = \sum_i (a_i)^2 + \sum_{i=1}^{n} c_{a,i} q_i \geq 0 \qquad b = \sum_i (b_i)^2 + \sum_{i=1}^{n} c_{b,i} q_i \geq 0$$

Just multiplying and adding gives

$$va + wb = \sum_i v(a_i)^2 + \sum_i w(b_i)^2 + \sum_{i=1}^{n} (vc_{a,i} + wc_{b,i}) q_i \geq 0$$

The case 4 is similar - we interpret strict equality as having $l = 0$ in Equation (1.4). In this case the degree is only the maximum of the two premises.

Notation We work with n variables $\mathbf{x} := x_1, \ldots, x_n$. Given a vector $\alpha \in \mathbb{N}^n$ ($\mathbb{N}$ including 0), we define $\mathbf{x}^\alpha := x_1^{\alpha_1} \ldots x_n^{\alpha_n}$, and in this context we call α a *degree vector*. The degree $|\alpha|$ of x^α is the sum of the coordinates $\sum_{i=1}^{n} \alpha_i$. Polynomials $p \in \mathbb{R}[\mathbf{x}]$ will be expressed as $\sum_\alpha p_\alpha \mathbf{x}^\alpha$, i.e. p_α is the coefficient of the monomial $\mathbf{x}^\alpha$ in p.

If q is some vector, by $\nu_d(q)$ we mean the vector consisting of all products of the coordinates of q of degree up to d. $s(d)$ represents the dimension of this vector and is calculated as $\binom{n+d-1}{d}$.

For any $n \in \mathbb{N}$, $[n]$ denotes the set $\{1, \ldots, n\}$.

For matrices A and B we use the notation $A \succeq 0$ to mean A is positive semidefinite and $A \succeq B$ to mean $A - B \succeq 0$. We will use the Frobenius inner product of two $m \times n$ matrices $\langle A, B \rangle := \sum_{i=1}^{m} \sum_{i=j}^{n} A_{ij} B_{ij}$, i.e. the dot product of the matrices laid out as vectors.

Polynomial optimization as a search for a probability measure A polynomial optimization problem (POP) is a problem of the form

$$\begin{aligned} \min_{x} \quad & o(\mathbf{x}) \\ \text{s.t.} \quad & g_i(\mathbf{x}) \geq 0 \qquad \forall i \in [c] \end{aligned} \tag{1.5}$$

for a collection of polynomials $o, g_i, \ldots g_c$. It is very hard in general - for example, the constraints $x_i^2 - x_i \geq 0, x_i - x_i^2 \geq 0$ force x_i to be binary.

We can turn any POP into a convex optimization problem by instead considering probability measures. Consider the problem

$$\begin{aligned} \min_{\mu} \quad & \int_K o(\mathbf{x}) d\mu \\ \text{s.t.} \quad & \int_K 1 d\mu = 1 \end{aligned} \tag{1.6}$$

Where K is the feasible region of (1.5) defined by the g_i. These problems have the same optimal value - any solution $\mathbf{x}^*$ to (1.5) can be converted into an equally valuable solution to (1.6) which is the probability measure with its entire mass concentrated at $\mathbf{x}^*$. In the reverse direction, the fact that μ is a probability measure means $\int_K p(x) d\mu$ is at least the minimum value of p on K and so the values coincide.

Moment relaxations We are given an infinite vector y indexed by degree vectors in $\mathbb{N}^n$. We are also told that y is the vector of *moments* under some hypothetical probability measure μ supported on a semialgebraic set of interest K – that is, $y_\alpha = \int_K \mathbf{x}^\alpha d\mu$ is a *mixed moment* of order $|\alpha|$, or the expectation of the monomial $\mathbf{x}^\alpha$ according to μ. What properties should we expect y to satisfy?

Firstly, as $\int_K 1 d\mu$ should be 1, so should y_0.

Secondly, for every $p \in \mathbb{R}[\mathbf{x}]$,

$$0 \leq \int_K p^2 d\mu = \int_K \sum_{\alpha,\beta} p_\alpha p_\beta \mathbf{x}^{\alpha+\beta} d\mu = \sum_{\alpha,\beta} p_\alpha p_\beta \int_K \mathbf{x}^{\alpha+\beta} d\mu.$$

Then the infinite *moment matrix* M_y indexed by degree vectors defined by $M_\mu(\alpha, \beta) := y_{\alpha+\beta}$ should be positive semidefinite.

Similarly, given any $g \in \mathbb{R}[\mathbf{x}]$ nonnegative on K and any $p \in \mathbb{R}[\mathbf{x}]$

$$0 \leq \int_K gp^2 d\mu = \int_K \left(\sum_\alpha g_\alpha \mathbf{x}^\alpha \right) \left(\sum_{\alpha,\beta} p_\alpha p_\beta \mathbf{x}^{\alpha+\beta} \right) d\mu$$
$$= \sum_{\alpha,\beta} p_\alpha p_\beta \sum_\gamma g_\gamma \int_K \mathbf{x}^{\alpha+\beta+\gamma} d\mu.$$

Therefore the *localizing matrix* for g defined as $M_{y,g}(\alpha,\beta) := \sum_\gamma g_\gamma y_{\alpha+\beta+\gamma}$ should also be positive semidefinite.

The *Lasserre relaxation*, defined imminently, is simply the semidefinite program got by making truncated versions of these demands.

Definition 1.8. Let d be the maximum degree of the o, g_is in (1.5), and let $t \geq d$. The level-t Lasserre relaxation of the equivalent problem (1.5) is

$$\begin{aligned} \min_{y \in \mathbb{R}^{\mathbb{N}^{2t}}} \quad & y \cdot o \\ \text{s.t.} \quad & y_0 = 1 \\ & (M_y|2t) \succeq 0 \\ & (M_{y,g_i}|2t) \succeq 0 \qquad \forall i \in [c] \end{aligned}$$

where $(M|x)$ is the finite principal submatrix of the infinite matrix M indexed by degree vectors of degree at most x.

(This area suffers from an unfortunate lack of standardization in terminology - the terms Sum Of Squares, Positivstellensatz, and Lasserre have all been used almost interchangeably by different authors. We have defined our versions of SOS and Positivstellensatz above, and for us Lasserre refers to the optimization program just defined. The relation between Lasserre and the other two follows next.)

Semidefinite programming and duality. The standard primal and dual forms for semidefinite programs are

$$\begin{aligned} \max \quad & \langle C, X \rangle \\ \text{s.t.} \quad & \langle A_i, X \rangle = b_i \qquad i \in [m] \\ & X \succeq 0 \end{aligned} \tag{SDP-P}$$

$$\begin{aligned} \min \quad & b \cdot y \\ \text{s.t.} \quad & \sum_{i=0}^{m} y_i A_i \succeq C \end{aligned} \tag{SDP-D}$$

In SDP-P, the decision variable is a matrix X, SDP-D, the vector y, and b is the m dimensional vector of the b_i.
We will rephrase Definition 1.8 as a SDP in the dual form. To do so we slice up the moment matrices by using the symmetric matrices indexed by degree vectors

$$(B_\kappa)_{\alpha,\beta} = \begin{cases} y_k & \text{if } \alpha+\beta=\kappa \\ 0 & \text{otherwise} \end{cases} \qquad (B_{\kappa,g})_{\alpha,\beta} = \begin{cases} g_\kappa & \text{if } \alpha+\beta\leq\kappa \\ 0 & \text{otherwise} \end{cases}$$

This gives us

$$M_y = \sum_\kappa B_\kappa \qquad M_{y,g_i} = \sum_\kappa B_{\kappa,g_i}$$

We now rewrite the problem in Definition 1.8 as

$$\begin{aligned} \min_{y\in\mathbb{R}^{\mathbb{N}^{2t}}} \quad & y\cdot o \\ \text{s.t.} \quad & \sum_{|\alpha|\leq 2t,\alpha\neq 0} y_\alpha B_\alpha \succeq -B_\varnothing \\ & \sum \sum_{|\alpha|\leq 2t,\alpha\neq 0} y_\alpha B_{\alpha,g_i} \succeq -B_{\varnothing,g_i} \qquad \forall i\in[c]. \end{aligned} \tag{1.7}$$

We can combine the two linear matrix inequalities into one by using block diagonal matrices, bringing it into the form (SDP-D). In this case we can split the primal decision variable X into Y and Z_i, Y and Z_i corresponding to the diagonal blocks of the coefficient matrices. In this case the primal problem is

$$\begin{aligned} \max \quad & \langle -B_\varnothing, Y\rangle + \sum_{i=0}^m \langle -B_{\varnothing,g_i}, Z_i\rangle \\ (\text{i.e. } \min \quad & Y_{1,1} + \sum_{i=0}^m (g_{i,\varnothing}(Z_i)_{1,1})) \\ \text{s.t.} \quad & \langle X, B_\alpha\rangle + \sum_{i=0}^m \langle Z_i, B_{\alpha,g_i}\rangle = o_\alpha \qquad |\alpha|\leq 2t, \alpha\neq 0 \\ & Y, Z_i \succeq 0 \end{aligned} \tag{1.8}$$

Interpretation of the primal

Theorem 1.4. *A degree $2d$ polynomial p is a Sum Of Squares (is SOS) if and only if there exists a positive semidefinite $k \times s(d)$ matrix C such that $p(x) = v_d(x)^\top C v_d(x)$.*

Proof. ($\Leftarrow$) The positive semidefiniteness of C guarantees a square root S such that $S^\top S = C$. So

$$p(x) = v_d(x)^\top S^\top S v_d(x) = (S v_d(x)) \cdot (S v_d(x)).$$

The ith coordinate of the vector $S v_d(x)$ is $\sum_{|\alpha| \leq d} S_{i\alpha} x^\alpha$ and so the right hand side is SOS.

The reverse is very similar. □

Definition 1.9. The *quadratic module* $Q(G)$ generated by a set of polynomials $G = \{g_1, \ldots, g_m\}$ is the set of all polynomials of the form

$$\sigma_0 + \sum_{i=1}^{m} \sigma_i^2 g_i$$

where the σ_i are all SOS.

Any $p \in Q(G)$ is non-negative on the semialgebraic set satisfying $g_1 \geq 0, \ldots, g_m \geq 0$. This leads to an alternate relaxation of (1.5):

$$\begin{aligned} \max_{\lambda \in \mathbb{R}} \quad & \lambda \\ \text{s.t.} \quad & o(x) - \lambda \in Q(G). \end{aligned} \tag{1.9}$$

We want to solve this relaxation with a semidefinite program. To do this, we introduce positive semidefinite matrices $Y, Z_i (i \in [m])$ with the goal of having

$$o(x) - \lambda = v_d(x)^\top Y v_d(x) + \sum_{i=1}^{m} \left(v_d(x)^\top Z_i v_d(x) g_i \right)$$

We write this as a semidefinite program by comparing the coefficients on each side of the equality using the matrices B_κ defined above. This gives us exactly the primal program (1.8) - that is, the Lasserre relaxation is dual to the SDP computing SOS certificates!

SDPS do not generally exhibit strong duality. In [50] the following is proven:

Theorem 1.5 ([50], Theorem 1)**.** *If the POP (1.5) contains a ball constraint $R - \sum_{x \in V} x^2 \geq 0$ with $R > 0$, then the dual relaxations (1.8), (1.9) exhibit strong duality. That is, if both are feasible, both share the same optimum value, and if one is unbounded, the other is infeasible.*

As we will focus on binary optimisation problems (containing the constraints $x^2 = x$ or $x^2 = 1$ for every variable x) we can without loss of generality assume all of our problems contain the ball constraint $f(x) = n - \sum_{x \in V} x^2 \geq 0$ and therefore these programs are strongly dual.

Pseudodistributions. We have just seen that the Lasserre program is dual to the program computing SOS certificates. The feasibility of the Lasserre program at a certain degree can then be used to certify nonexistence of SOS certificates - and this can be seen more directly, without resorting to SDPS.

We are about to introduce *pseudodistributions*, the main device used to prove lower bounds for SOS. A degree-d pseudodistribution is just a feasible point for the d-th level Lasserre relaxation, and the fact (proven just above) that this relaxation is dual to the semidefinite program calculating SOS certificates already tells us that there can be no SOS refutation of degree-d if a degree-d pseudodistribution exists.
But for the sake of intuition, we point out that a pseudodistribution 'acts like' it is distributed over actual solutions of a polynomial optimisation problem, in a manner made precise below in Theorem 1.8. The higher the degree, the better the acting, and then the harder any refutation has to work in order to show that there are no actual solutions over which this pseudodistribution is distributed across. (This is where the term moment comes from - the value of this pseudodistribution for a monomial is pretending to be the expectation of that monomial according to some fictitious distribution.) We want to prove the existence of pseudodistributions for as high a degree as possible.

Definition 1.10. A *degree d pseudodistribution* L for a set of polynomial axioms $h_1 \geq 0, \ldots, h_x \geq 0, f_1 = 0, \ldots, f_y = 0$, all in $\mathbb{R}[V]$ and assumed to contain the boolean constraints $x^2 = x$ for all $x \in V$, is a linear mapping $\mathbb{R}[V] \to \mathbb{R}$ such that:

1. The *moment matrix* $M(\alpha, \beta) := L[\alpha \cup \beta]$ indexed by pairs of monomials α, β, is PSD, wherever the degree of α and β is restricted to be at most $d/2$,

2. for every axiom of the form $f = 0$ and any polynomial p with degree bounded by $d - deg(f)$ we have $L[pf] = 0$,

3. for every axiom of the form $q \geq 0$, its *localizing matrix*

$$\mathcal{M}_q(\alpha, \beta) := \sum_{\gamma \in \mathcal{P}(V, \leq d)} q_\gamma L[\alpha \cup \beta \cup \gamma],$$

is PSD, whenever $|\alpha|$ and $|\beta|$ are at most $(d - deg(q))/2$, where q_γ is the coefficient of q in front of the monomial γ, and

4. $L[1] = 1$.

(Due to the boolean constraints multiplication is idempotent, whence the union vs addition, and sets vs monomials.) A degree d pseudodistribution L immediately gives a degree d lower bound on a SOS refutation. This is because if we apply L to both sides of Equation (1.4) we find -1 on the left hand side but

1. If $deg(q^2) \leq d$ then

$$L[q^2] = L\left[\left(\sum_{\alpha \in \mathcal{P}(V, \leq d/2)} q_\alpha \alpha\right)^2\right] = \sum_{\alpha,\beta \in \mathcal{P}(V, \leq d/2)} q_\alpha q_\beta L[\alpha \cup \beta] = q^\top \mathcal{M} q \geq 0,$$

   with the last inequality coming from the PSDness of $\mathcal{M}$,

2. for every axiom of the form $f = 0$ and any polynomial p with degree bounded by $d - deg(f)$ we have $L[pf] = 0$,

3. for every axiom of the form $q \geq 0$ and any polynomial p such that $deg(p^2 q) \leq d$, we have

$$\begin{aligned} &L[p^2 q] \\ &= L\left[\left(\sum_{\alpha,\beta \in \mathcal{P}(V, \leq ((d - deg(q))/2)} p_\alpha p_\beta \, (\alpha \cup \beta)\right)\left(\sum_{\gamma \in \mathcal{P}(V, \leq deg(q))} q_\gamma \gamma\right)\right] \\ &= \sum_{\alpha,\beta \in \mathcal{P}(V, \leq ((d - deg(q))/2)} p_\alpha p_\beta \left(\sum_{\gamma \in \mathcal{P}(V, \leq deg(q))} q_\gamma L[\alpha \cup \beta \cup \gamma]\right) = p^\top \mathcal{M}_q p \geq 0, \end{aligned}$$

   with the last inequality coming from the PSDness of $\mathcal{M}_q$.

SOS relaxations have found many successful applications. To name a few:

1. The Goemans-Williamson algorithm is a SDP giving an approximate solution to the max-cut problem. It can be rephrased as the degree 2 SOS SDP. This algorithm is optimal assuming the Unique Games Conjecture holds. For more, see, for example, [68].

2. Given some graph G, the Lovasz theta function $\theta(G)$ is the solution to an SDP given by Lovasz in [63], and satisfies

$$\omega(G) \leq \theta(G) \leq \chi(G)$$

where $\omega(G)$ is the size of the largest clique in G and $\chi(G)$ is the smallest number of colours used in any proper colouring of G. The class of *perfect graphs* are the graphs where ω and χ coincide, and for these graphs $\theta(G)$ is just the clique/chromatic number of the graph. Again, it turns out to be captured by the degree 2 SOS SDP [3], therefore SOS gives an efficient algorithm to solve the clique/k-colouring problem for perfect graphs. In fact, a long-standing open problem in graph theory is to find a more combinatorial algorithm for computing these numbers in perfect graphs that do not resort to semidefinite programming.

3. Sparse principal component analysis can be carried out using semidefinite programming. Again, it turns out that the basic SDP approach, which requires k^2 samples, is equivalent to the degree 2 SOS SDP.

4. An extremely successful application of SOS is given in [61]. There it is proven that *any* polynomial size SDP that approximates some max-CSP is equivalent in power to some constant degree SOS SDP.

Despite its power, there are some existing lower bounds:

1. **3-XOR.** Refuting a randomly generated 3-XOR instance (which is unsatisfiable w.h.p) is hard for SOS. A linear lower bound is given in [42]. Note that this is doable with Gaussian elimination, in polynomial time!

2. **Planted clique.** A series of works, culminating in the following result:

   Theorem 1.6 ([4], Theorem 1.1, informal.)**.** *Let $d \in o(\log n)$. Then, with high probability, the degree d SOS for max-clique will return an optimal value of $\Omega(\sqrt{n})$ for an Erdős-Rényi random graph with edge probability $1/2$ - although such a graph only has a clique of size more than $2\log n$ with probability exponentially diminshing in n.*

3. The so-called '**Knapsack**' contradiction, which is just that if we have some boolean variables $x_1, \ldots, x_n$ and some non-integral r, we can never have $\sum_{i=1}^{n} x_i = r$. In [41] the following linear lower bound is given:

   Theorem 1.7 ([41], Proposition 1)**.** *If $r \leq n/2$ then the Knapsack problem only has SOS refutations of degree $\Omega(r)$.*

4. **Sparse PCA** [65]. Here it is proven that the degree 4 SOS algorithm for sparse PCA does not require many fewer samples than the degree 2 SOS algorithm.

5. **The Least Ordering Principle** [71]. Here a $O(\sqrt{n}\log n)$ upper bound, and a non-constant lower bound, is given for the Least Ordering Principle.

The existence of a degree $O(n)$ SOS refutation for some simultaneously unsatisfiable polynomials, that is, the completeness of SOS as a proof system, can be shown in different ways. For example, it can follow from the strong duality of theorem 1.5. We state here instead a result given in [58, 80] that is stronger than completeness:

Theorem 1.8 ([80], Theorem 9)**.** *Suppose you are given a point x feasible for the tth level Lasserre relaxation. Then, for any $S \subseteq [n]$ with $|S| \leq t$, x is a convex combination of points feasible for the $(t - |S|)$th Lasserre relaxation which are all $0/1$ on the indices S.*

1.2.4 Sherali-Adams

The Sherali-Adams (SA) proof system was introduced first by Sherali and Adams [83] as a means of solving ILPs, and then later considered as a propositional refutation system in [24]. Since then it has been considered as a refutation system in the further works [2, 29].

Definition 1.11. Let $A, B \subseteq V$ with $A \cap B = \varnothing$ and $|A \cup B| \leq t$ for some $t \in \mathbb{N}$. Denote by $P_{A,B}$ the product $\prod_{x_i \in A} x_i \prod_{x_j \in B}(1 - x_j)$. A *Sherali-Adams (SA) refutation* of $\{f_1 = 0, \ldots, f_l = 0\, h_1 \geq 0, \ldots, h_r \geq 0\}$ is an identity of the form

$$\sum (a_{I,J} P_{A_I,B_J} h_i) + \sum (p_i f_i) + \sum q_i (x_i^2 - x) = -1 \tag{1.10}$$

where $a_{I,J} \in \mathbb{R}$ are nonnegative and the p_i, q_i are arbitrary polynomials. The degree of this proof is the maximum degree of any of the $P_{A_I,B_J}, p_i f_i, q_i(x_i^2 - x)$.

Similarly as for SOS, the coefficients in Equation (1.10) can be found by linear programming. When proving lower bounds (which we do in Chapter 4) we will actually show that the dual of this LP is feasible - this is analogous to finding a pseudodistribution for SOS. We describe now the dual LP.

Let C be some CNF using variables $v_1, \ldots, v_m$. We generate the following LP over the $2m$ variables $Z_{v_i}, Z_{\neg v_i}$, $1 \leq i \leq m$. For each clause $(l_1 \vee \ldots \vee l_t)$ in C we have the constraining inequality

$$Z_{l_1} + \ldots + Z_{l_t} \geq 1.$$

We also have, for each $\lambda \in [m]$, the equalities of negation

$$Z_{v_\lambda} + Z_{\neg v_\lambda} = 1 \tag{1.11}$$

together with the bounding inequalities

$$0 \leq Z_{v_\lambda} \leq 1 \quad \text{and} \quad 0 \leq Z_{\neg v_\lambda} \leq 1.$$

Let $\mathcal{P}_0^C$ be the polytope specified by these constraints on the real numbers. We noted above that this polytope contains integral points if and only if the formula C is satisfiable. However even if C is unsatisfiable, $\mathcal{P}_0^C$ may be non-empty; in fact, if F is a contradiction that does not admit refutation by unit clause propagation, this is the case (we may use unit clause propagation to assign $0-1$ values to some variables, thereafter assigning $1/2$ to those variables remaining). Note that it follows that any unsatisfiable Horn CNF C (i.e., where each clause contains at most one positive variable) has SA rank 0, since C must then admit refutation by unit clause propagation (which may be used to demonstrate $\mathcal{P}_0^C$ empty).

Sherali-Adams provides a static refutation method that takes the polytope $\mathcal{P}_0^C$ and *r-lifts* it to another polytope $\mathcal{P}_r^C$ in $\sum_{\lambda=0}^{r+1} \binom{2m}{\lambda}$ dimensions. Specifically, the variables involved in defining the polytope $\mathcal{P}_r^C$ are $Z_{l_1 \wedge \ldots \wedge l_{r+1}}$ ($l_1, \ldots, l_{r+1}$ literals of C) and $Z_\emptyset$. We say that the term $Z_{l_1 \wedge \ldots \wedge l_{r+1}}$ has *rank* r. Note that we accept commutativity and idempotence of the $\wedge$-operator, e.g. $Z_{l_1 \wedge l_2} = Z_{l_2 \wedge l_1}$ and $Z_{l_1 \wedge l_1} = Z_{l_1}$. Also $\emptyset$ represents the empty conjunct, which is boolean true; hence we set $Z_\emptyset := 1$. For literals $l_1, \ldots, l_t$, s.t. $(l_1 \vee \ldots \vee l_t)$ is a clause of C, we have the constraining inequalities

$$Z_{l_1 \wedge D} + \ldots + Z_{l_t \wedge D} \geq Z_D,$$

for D any conjunction of at most r literals of C. We also have, for each $\lambda \in [m]$ and D any conjunction of at most r literals, the equalities of negation

$$Z_{v_\lambda \wedge D} + Z_{\neg v_\lambda \wedge D} = Z_D$$

together with the bounding inequalities

$$0 \leq Z_{v_\lambda \wedge D} \leq Z_D \quad \text{and} \quad 0 \leq Z_{\neg v_\lambda \wedge D} \leq Z_D.$$

For $r' \leq r$, the defining inequalities of $\mathcal{P}_{r'}^C$ are consequent on those of $\mathcal{P}_r^C$. Equivalently, any solution to the inequalities of $\mathcal{P}_r^C$ gives rise to solutions of the inequalities of $\mathcal{P}_{r'}^C$, when projected on to its variables. If D' is a conjunction of r' literals, then $Z_{D \wedge D'} \leq Z_D$ follows by transitivity from r' instances of the bounding inequalities defined last. We refer to the property $Z_{D \wedge D'} \leq Z_D$ as *monotonicity*. Finally we note that $Z_{v \wedge \neg v} = 0$ holds in $\mathcal{P}_1^C$ and follows from a single lift of an equality of negation.

The SA *rank* of the polytope $\mathcal{P}_0^C$ is the minimal i such that $\mathcal{P}_i^C$ is empty. Thus, the

notation rank is overloaded in a consistent way, since $\mathcal{P}_i^C$ is specified by inequalities in variables of rank at most i. The largest r for which $\mathcal{P}_r^C$ need be considered is $2m-1$, since beyond that there are no new literals to lift by. Even that is somewhat further than necessary, largely because, if the conjunction D contains both a variable and its negation, it may be seen from the equalities of negation that $Z_D = 0$. In fact, it follows from [58] that the SA rank of $\mathcal{P}_0^C$ is always $\leq m-1$ (for a contradiction C).

The number of defining inequalities of the polytope $\mathcal{P}_r^C$ is exponential in r; hence a naive measure of SA size would see it grow more than exponentially in rank. However, not all of the inequalities may be needed to specify the empty polytope. We therefore define the SA *size* of the polytope $\mathcal{P}_0^C$ to be the size (of an encoding) of a minimal subset of the inequalities in $\mathcal{P}_{2m}^C$ needed to specify the empty polytope.

1.2.5 Cutting Planes

Cutting Planes was introduced by Gomory in [39, 40] as a method of solving ILPs, and formalised later as a proof system in [18, 21].

Definition 1.12. A *linear integer inequality* in the variables $x_1, \ldots, x_n$ is an expression of the form $\sum_{i=1}^n a_i x_i \geq b$, where each a_i and b are integral. A set of such inequalities is said to be *unsatisfiable* if there are no 0/1 assignments to the x variables satisfying every inequality simultaneously.

Note that we reserve the term infeasible, in contrast to unsatisfiable, for (real or rational) linear programs.

Definition 1.13. The *Cutting Planes* (CP) proof system is equipped with boolean axioms and two inference rules:

$$\text{Boolean axioms:} \qquad 0 \leq x \leq 1 \qquad \text{for any variable } x$$

$$\text{Linear combination:} \qquad \mathbf{a}\cdot\mathbf{x} \geq c, \mathbf{b}\cdot\mathbf{x} \geq d \implies \alpha\mathbf{a}\cdot\mathbf{x} + \beta\mathbf{b}\cdot\mathbf{x} \geq \alpha c + \beta d$$

$$\text{Rounding:} \qquad \alpha\mathbf{a}\cdot\mathbf{x} \geq b \implies \mathbf{a}\cdot\mathbf{x} \geq \lceil b/\alpha \rceil$$

where $\alpha, \beta, b \in \mathbb{Z}^+$ and $\mathbf{a}, \mathbf{b} \in \mathbb{Z}^n$. A CP refutation of some unsatisfiable set of integer linear inequalities $\mathcal{F}$ is a derivation of $0 \geq 1$ by the aforementioned inference rules from the inequalities in $\mathcal{F}$.

A CP refutation is *treelike* if the directed acyclic graph underlying the proof is a tree. The *length* of a CP refutation is the number of inequalities in the sequence. The *depth*

is the length of the longest path from the root to a leaf (sink) in the graph. The *rank* of a CP proof is the maximal number of rounding rules used in a path of the proof graph. The *size* of a CP refutation is the bit size required to represent all the inequalities in the proof.

It is sometimes convenient, as was the case in e.g. [15], to think about Cutting Planes from a 'dual' aspect, in terms of the application of a closure operator on some polytope.

Definition 1.14. Given some polytope $P \subseteq \mathbb{R}^n$, we let P' be the set of points in P that 'survive all cutting planes':

$$P' := \{x \in P \ : \ \mathbf{a} \cdot x \geq \lceil b \rceil \text{ for all } \mathbf{a} \in \mathbb{Z}^n, b \in \mathbb{R} \text{ satisfying } \mathbf{a} \cdot y \geq b \quad \forall y \in P\}.$$

Given the infinite quantification over the $\mathbf{a} \in \mathbb{Z}^n$, it is certainly not obvious that P' should be a polytope if P is. Schrijver shows in [81] that if P is a rational polytope, in the sense that it is specified by finitely many linear inequalities with rational coefficients, then so is P'. This result was generalised to the irrational (real) case by Dunkel in her PhD thesis [32].
Now we give an equivalent definition of Cutting Planes rank:

Definition 1.15. Let P be a polytope. Letting $P^{(0)} := P$, $P^{(i+1)} := P^{(i)\prime}$, and P_I be the convex hull of integral points in P, the *rank* of P is the $r \in \mathbb{N}$ with $P^{(r)} = P_I$.

That this measure is even well defined, in the sense that this operator converges onto the integer hull after finitely many cuts, was shown in Chvatal in [19]. We state here a quantitative completeness given in [33] most relevant to the polytopes appearing in proof complexity:

Theorem 1.9 ([33], Section 3)**.** *If $P \subseteq \mathbb{R}^n$ is a polytope contained in the unit hypercube $[0,1]^n$, then its CP rank is at most $O(n^2 \log n)$.*

1.2.6 Stabbing Planes

Stabbing Planes, introduced recently in [7] by Beame, Fleming, Impagliazzo, Kolokolova, Pankratov, Pitassi and Robere, is another algorithm for refuting unsatisfiable ILPs.

DPLL, perhaps the most famous approach to SAT solving, searches for an assignment satisfying some CNF by choosing some variable x and then recursively solving the two problems gotten for each possible boolean setting of x. Of course, turning some DPLL execution trace upside-down gives you a (treelike) Resolution refutation, so long as your instance was indeed unsatisfiable. Stabbing Planes might be described as a geometric

edition of DPLL - instead of asking about the boolean truth value of some variable and eventually proving some CNF unsatisfiable, we ask if some *integer sum* $\sum_i z_i v_i$, $z_i \in \mathbb{Z}$, of some *continuous real variables* v_i is at least some $b \in \mathbb{Z}$ or at most $b-1$, and eventually prove some polytope integer free. In [7] this description is made precise: SP is polynomially equivalent to the proof system treelike Res(CP), introduced by Krajíček in [54], where clauses are now disjunctions of linear inequalities.

Definition 1.16. Fix some variables $x_1, \ldots, x_n$. A *Stabbing Planes* (SP) refutation of an unsatisfiable set of integer linear inequalities $\mathcal{F}$ is a binary tree $\mathcal{T}$, with each node labeled by a *query* $(\mathbf{a}, b)$ with $\mathbf{a} \in \mathbb{Z}^n, b \in \mathbb{Z}$. Out of each node we have an edge labeled with $\mathbf{a} \cdot x \geq b$ and the other labeled with its integer negation $\mathbf{a} \cdot x \leq b-1$. Each leaf ℓ is labeled with an infeasible LP system P_ℓ made by a nonnegative linear combination of inequalities from $\mathcal{F}$ and the inequalities labelling the edges on the path from the root of $\mathcal{T}$ to the leaf ℓ.

The *length* of a SP refutation is the number of queries in the proof tree. The *depth* of a SP refutation $\mathcal{T}$ is the longest root-to-leaf path in $\mathcal{T}$. The size (respectively depth) of refuting $\mathcal{F}$ in SP is the *minimum* size (respectively depth) over all SP refutations of $\mathcal{F}$.

Note that here we do not care about the complexity of actually witnessing the infeasibility of the P_ℓ, which is itself (anything up to a) polynomial. Often a proof system (such as Resolution) will come with trivial and immediate linear size lower bound because, for example, every axiom should be downloaded at least once somewhere, but even if the input is minimally unsatisfiable this is not necessarily true for SP. For example, a minimally unsatisfiable Horn-CNF will have a size 1 SP refutation but only linearly sized SA refutations (albeit 0 rank).

Definition 1.17. The *slab* corresponding to a query $Q = (\mathbf{a}, b)$ is the set $\mathsf{slab}(Q) = \{\mathbf{x} \in \mathbb{R}^n : b-1 < \mathbf{a} \cdot \mathbf{x} < b\}$ satisfying neither of the associated inequalities.

Since each leaf in a SP refutation is labelled by an infeasible LP, in this thesis we will often use the following geometric observation about SP refutations $\mathcal{T}$: the set of points in $\mathbb{R}^n$ that are not already ruled out by an axiom must all be ruled out by a query somewhere in $\mathcal{T}$. In particular this will be true for those points in $\mathbb{R}^n$ which satisfy a set of integer linear inequalities $\mathcal{F}$ and which we call *feasible points* for $\mathcal{F}$.

Lemma 1.2. *The slabs associated with a SP refutation must cover the feasible points of* $\mathcal{F}$. *That is,*

$$\{\mathbf{y} \in \mathbb{R}^n : \mathbf{a} \cdot \mathbf{y} \geq b \text{ for all } (\mathbf{a}, b) \in \mathcal{F}\} \subseteq \bigcup_{(\mathbf{a},b) \in \mathcal{T}} \{\mathbf{x} \in \mathbb{R}^n : b-1 < \mathbf{a} \cdot \mathbf{x} < b\}$$

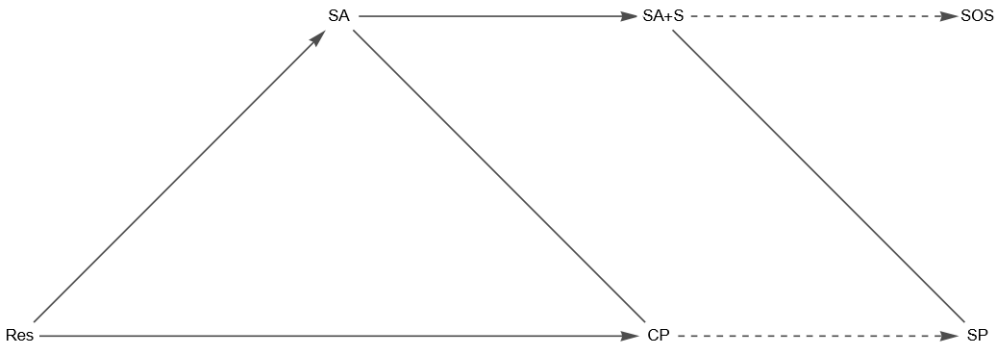

FIGURE 1.1: A diagram of the relative strengths of the proof systems discussed in this thesis. A solid directed edge indicates that the source is exponentially weaker than the sink. A dashed directed edge indicates that the source is simulated by the sink, but that no exponential separation is known. An undirected edge indicated incomparability.

1.3 Thesis outline

This thesis comes in three main chapters. In the first and last, we are mostly focused on the proof complexity of some particular proof system, and in the second, the complexity in general of principles given in FO language.

In the following Chapter 2 we investigate the proof complexity of the recently introduced Stabbing Planes proof system. The strongest lower bounds known for this system are 'indirect', in the sense of their appealing to simulations and to results in communication complexity. We introduce a number of methods of proving lower bounds, all of which have some geometric flavour, and all working directly against the proof system in question.

In the sequel Chapter 3, we make progress towards proving broad lower bounds for principles specified in FO logic. We start by using machinery from Chapter 2 to show that any FO principle with only infinite models requires polynomial size (and therefore logarithmic rank) to refute in Stabbing Planes. In doing so, we generalise a subset of the results from Chapter 2. We then turn to a stronger proof system, Sum Of Squares, and while we do not achieve the same sort of lower bound, we provide some results we believe are insightful.

In Chapter 4, we investigate the the effect of choosing alternative encodings (as propositional formulae) for some FO principles which at this point we now find familiar. These encodings differ mainly in how they deal with the translation of existential demands. We find some results that, at least to the author, seem counter-intuitive.

The thesis then concludes by indicating some natural further directions.

Chapter 2 was published in paper form as 'Depth Lower Bounds in Stabbing Planes for Combinatorial Principles' in STACS [25], and is joint work with the authors there.

A prior version of Chapter 4 was published in paper form as 'Sherali-Adams and the Binary Encoding of Combinatorial Principles' in LATIN [27] and is likewise joint work with the authors there. Chapter 3 remains unpublished.

Chapter 2

Stabbing Planes

Stabbing Planes (also known more commonly outside of proof complexity as Branch and Cut) is a proof system introduced very recently which, informally speaking, extends the DPLL method by branching on integer linear inequalities instead of single variables. The techniques known so far to prove size and depth lower bounds for Stabbing Planes are generalizations of those used for the Cutting Planes proof system established via communication complexity arguments. As such they work for composed versions of combinatorial statements. Rank lower bounds for Cutting Planes are also obtained by geometric arguments called *protection lemmas*.

In this work we introduce two new geometric approaches to prove size/depth lower bounds in Stabbing Planes working for a wide range of principles: (1) the *antichain method*, relying on *Sperner's Theorem* and (2) the *covering method* which uses results on *essential coverings* of the boolean cube by linear polynomials, which in turn relies on *Alon's combinatorial Nullenstellensatz*.

We demonstrate their use on classes of combinatorial principles such as the Pigeonhole Principle, Tseitin Priniples, and the Linear Ordering Principle. By the first method we prove almost linear size lower bounds and optimal logarithmic depth lower bounds for the Pigeonhole Principle, analogous lower bounds for the Tseitin Principles over the complete graph, and for the Linear Ordering Principle. Via the covering method we obtain a superlinear size lower bound and a logarithmic depth lower bound for Stabbing Planes proof of Tseitin Principles over a grid graph.

Finally, as a specific result, we prove that for any graph G, any SP refutation of a Tseitin Principle for G has length at least the so called *circuit rank* of G, which, if G has c connected components, is $E(G) - V(G) + c$.

2.1 Introduction

Finding a satisfying assignment for a propositional formula (SAT) is a central component for many computationally hard problems. Despite being older than 50 years and exponential time in the worst-case, the DPLL algorithm [30, 31] is the core of essentially all high performance modern SAT-solvers. DPLL is a recursive boolean method: at each recursive call, one variable x of the formula $\mathcal{F}$ is chosen and the search recursively branches into the two cases obtained by setting x respectively to 1 and 0 in $\mathcal{F}$. The worst cases for DPLL are unsatisfiable instances, as there it must explore every possible assignment, and it is well-known that the execution trace of the DPLL algorithm running on an unsatisfiable formula $\mathcal{F}$ is nothing more than a treelike refutation of $\mathcal{F}$ in the proof system of Resolution [79]. Since SAT can be viewed as an optimization problem the question whether Integer Linear Programming (ILP) can be made feasible for satisfiability testing received a lot of attention and is considered among the most challenging problems in local search [51, 82]. One proof system capturing ILP approaches to SAT is *Cutting Planes*, a system whose main rule implements the *rounding* (or *Chvátal cut*) approach to ILP. Cutting planes works with integer linear inequalities of the form $\mathbf{ax} \leq b$, with $\mathbf{a}, b$ integers, and, like Resolution, is a sound and complete refutational proof system for CNF formulas: indeed a clause $C = (x_1 \vee \ldots \vee x_r \vee \neg y_1 \vee \ldots \vee \neg y_s)$ can be written as the integer inequality $\mathbf{y} - \mathbf{x} \leq s - 1$.

Beame et al. [7] extended the idea of DPLL to a more general proof strategy based on ILP. Instead of branching only on a variable as in Resolution, in this method one considers a pair $(\mathbf{a}, b)$, with $\mathbf{a} \in \mathbb{Z}^n$ and $b \in \mathbb{Z}$, and branches limiting the search to the two half-planes: $\mathbf{ax} \leq b - 1$ and $\mathbf{ax} \geq b$. A *path* terminates when the LP defined by the inequalities in $\mathcal{F}$ and those forming the path is infeasible. This method can be made into a refutational treelike proof system for unsatisfiable CNFs called Stabbing planes (SP) ([7]) and it turns out that it is polynomially equivalent to the treelike version of Res(CP), a much older proof system introduced by Krajíček [54] where clauses are disjunctions of linear inequalities. The Stabbing Planes proof system is defined precisely in Section 1.2.6.

In this work we consider the complexity of proofs in SP focusing on the *length*, i.e. the number of queries in the proof, and the *depth* (called *rank* in [7]), i.e. the length of the longest path in the proof tree. Note that the length metric is at least as strong as the *size* metric (appearing in [35]), which is the bit size of all the coefficients appearing in the proof. The results in this chapter are stated for size, however, note that they are actually (stronger) length lower bounds.

2.1.1 Previous works and motivations

After its introduction as a proof system in the work [7], *Stabbing Planes* received great attention. The quasipolynomial upper bound for the size of refuting Tseitin contradictions in SP given in [7] was surprisingly extended to CP by of Dadush and Tiwari [22] refuting a long-standing conjecture. Recently in [35], Fleming, Göös, Impagliazzo, Pitassi, Robere, Tan and Wigderson developed on the ideas given in [7] making important progress on the question whether all Stabbing Planes proofs can be somehow efficiently simulated by Cutting Planes.

Significant lower bounds for size can be obtained in SP, but in a limited way, using modern developments of a technique for CP based on communication complexity of search problems introduced by Impagliazzo, Pitassi, Urquhart in [49]: in [7] it is proven that size S and depth D SP refutations imply treelike $\mathsf{Res}(\mathrm{CP})$ proofs of size $O(S)$ and width $O(D)$; Kojevnikov [53], improving the *interpolation method* introduced for $\mathsf{Res}(\mathrm{CP})$ by Krajíček [54], gave exponential lower bounds for treelike $\mathsf{Res}(\mathrm{CP})$ when the width of the clauses (i.e. the number of linear inequalities in a clause) is bounded by $o(n/\log n)$. Hence these lower bounds are applicable only to very specific classes of formulas (whose hardness comes from boolean circuit hardness) and only to SP refutations of low depth.

Nevertheless SP appears to be a strong proof system. As noted in Section 1.2.6, the condition terminating a path in a proof is not a trivial contradiction like in Resolution, but is the infeasibility of an LP, which is only a polynomial time verifiable condition. Hence linear size SP proofs might be already a strong class of SP proofs, since they can hide a polynomial growth into one final node whence to run the verification of the terminating condition.

Rank and depth in CP and SP

It is known that, contrary to the case of other proof systems like Frege, neither CP nor SP proofs can be balanced (see [7]), in the sense that a depth-d proof can always be transformed into a size $2^{O(d)}$ proof. The depth of CP-proofs of a set of linear inequalities L is measured by the Chvátal rank of the associated polytope P. It is known that rank in CP and depth in SP are separated, in the sense that Tseitin Principles can be proved in depth $O(\log^2 n)$ depth in SP [7], but are known to require rank $\Theta(n)$ to be refuted in CP [15].

Rank lower bound techniques for Cutting Planes are essentially of two types. The main method is by reducing to the real communication complexity of certain search problem [49]. As such this method only works for classes of formulas *lifted* by certain gadgets

capturing specific boolean functions. A second class of methods have been developed for Cutting Planes, which lower bound the rank measures of a polytope. In this setting, lower bounds are typically proven using a geometric method called *protection lemmas* [15]. These methods were recently extended in [35] also to the case of Semantic Cutting Planes (a strengthened version of Cutting Planes, where from any two inequalities $A, B \geq 0$, you may infer any $C \geq 0$ that is a sound inference assuming 0/1 assignments). In principle this geometric method can be applied to any formula and not only to the lifted ones, furthermore for many formulas (such as the Tseitin Principles) it is known how to achieve $\Omega(n)$ rank lower bounds in CP via protection lemmas, while proving even $\omega(\log n)$ lower bounds via real communication complexity is impossible, due to a known folklore upper bound.

Lower bounds for depth in Stabbing Planes, proved in [7], are instead obtained only as a consequence of the real communication approach extended to Stabbing Planes. In this chapter we introduce several geometric approaches to prove depth lower bounds in SP.

Specifically the results we know at present relating SP and CP are:

1. SP polynomially simulates CP (Theorem 4.5 in [7]). Hence in particular the PHP^m_n can be refuted in SP by a proof of size $O(n^2)$ ([21]). Furthermore it can be refuted by a $O(\log n)$ depth proof since polynomial size CP proofs, by Theorem 4.4 in [7], can be balanced in SP [1].

2. Beame et al. in [7] proved the surprising result that the class of Tseitin contradictions $\mathsf{Ts}(G, \omega)$ over any graph G of maximum degree D, with an odd charging ω, can be refuted in SP in size quasipolynomial in $|G|$ and depth $O(\log^2 |G| + D)$.

Depth lower bounds for SP are proved in [7]:

1. a $\Omega(n/\log^2 n)$ lower bound for the formula $\mathsf{Ts}(G, w) \circ \mathsf{VER}^n$, composing $\mathsf{Ts}(G, \omega)$ (over an expander graph G) with the gadget function VER^n (see Theorem 5.7 in [7] for details); and

2. a $\Omega(\sqrt{n \log n})$ lower bound for the formula $\mathsf{Peb}(G) \circ \mathsf{IND}^n_l$ over $n^5 + n \log n$ variables obtained by lifting a pebbling formula $\mathsf{Peb}(G)$ over a graph with high pebbling number, with a *pointer function* gadget IND^n_l (see Theorem 5.5. in [7] for details).

[1] Another way of proving this result is using Theorem 4.8 in [7] stating that if there are length L and space S CP refutations of a set of linear integral inequalities, then there are depth $O(S \log L)$ SP refutations of the same set of linear integral inequalities; and then use the result in [37] (Theorem 5.1) that PHP^m_n has polynomial length and constant space CP refutations.

Similar to size, these depth lower bounds are applicable only to very specific classes of formulas. In fact they are obtained by extending to SP the technique introduced in [49, 55] for CP of reducing shallow proofs of a formula $\mathcal{F}$ to efficient *real* communication protocols computing a related search problem and then proving that such efficient protocols cannot exist.

Despite the fact that SP is at least as strong as CP, in SP the known lower bound techniques are derived from those of treelike CP. Hence finding other techniques to prove depth and size lower bounds for SP is important to understand its proof strength. For instance, unlike CP where we know tight $\Theta(\log n)$ rank bounds for the PHP_n^m [15, 76] and $\Omega(n)$ rank bounds for Tseitin contradictions [15], for SP no depth lower bound is at present known for purely combinatorial statements.

In this chapter we address such problems.

2.1.2 Contributions and techniques

The main motivation of this work is to study size and depth lower bounds in SP through new methods, possibly geometric. Differently from weaker systems like Resolution, except for the technique highlighted above and based on reducing to the communication complexity of search problems, we do not know of other methods to prove size and depth lower bounds in SP. In CP and Semantic CP instead geometrical methods based on protection lemmas were used to prove rank lower bounds in [15, 35].

Our first steps in this direction were to set up methods working for truly combinatorial (so, uncomposed and gadget-free) statements, like $\mathsf{Ts}(G, w)$ or PHP_n^m, which we know to be efficiently provable in SP, but on which we cannot use methods reducing to the complexity of boolean functions, like the ones based on communication complexity.

We present two new and fairly general methods for proving depth lower bounds in SP which in fact are the consequence of proving length lower bounds that do not depend on the bit-size of the coefficients.

As applications of our two methods we respectively prove:

1. An exponential separation between the rank in CP and the depth in SP, using a new counting principle which we introduce and that we call the *Simple Pigeonhole Principle* (SPHP). We prove that the SPHP has $O(1)$ rank in CP and requires $\Omega(\log n)$ depth in SP. Together with the results proving that Tseitin formulas requires $\Omega(n)$ rank lower bounds in CP ([15]) and $O(\log^2 n)$ upper bounds for the depth in SP ([7]), this proves an incomparability between the two measures.

2. An almost linear lower bound for the size of SP proofs of the PHP_n^m and for Tseitin $\mathsf{Ts}(G,\omega)$ contradictions over the complete graph. These lower bounds immediately give optimal $\Omega(\log n)$ lower bound for the depth of SP proofs of the corresponding principles.

3. A superlinear-in-n lower bound for the size of SP proofs of $\mathsf{Ts}(G,\omega)$, when G is a $n \times n$ grid graph H_n. In turn this implies an $\Omega(\log n)$ lower bound for the depth of SP proofs of $\mathsf{Ts}(H_n,\omega)$. Proofs of depth $O(\log^2 n)$ for $\mathsf{Ts}(H_n,\omega)$ are given in [7].

4. Finally we prove linear lower bound for the size and $O(\log n)$ lower bounds of the depth for the the Linear Ordering Principle LOP.

All of our results are derived from the following initial geometrical observation: let $\mathbb{S}$ be a space of *admissible points* in $\{0,1,1/2\}^n$ satisfying a given unsatisfiable system of integer linear inequalities $\mathcal{F}(x_1,\ldots,x_n)$. In a SP proof for $\mathcal{F}$, at each branch $q = (\mathbf{a}, b)$ the set of points in the $\mathsf{slab}(q) = \{\mathbf{s} \in \mathbb{S} : b-1 < \mathbf{ax} < b\}$ does not survive in $\mathbb{S}$. At the end of the proof on the leaves, where we have infeasible LP's, no point in $\mathbb{S}$ can survive the proof. So it is sufficient to find conditions such that, under the assumption that a proof of $\mathcal{F}$ is 'small', even one point of $\mathbb{S}$ survives the proof. In pursuing this approach we use two methods.

The *antichain method*. Here we use a well-known bound based on Sperner's Theorem [17, 87] to upper bound the number of points in the slabs where the set of non-zero coefficients is sufficiently large. Trading between the number of such slabs and the number of points ruled out from the space $\mathbb{S}$ of admissible points, we obtain the lower bound.

We initially present the method and the $\Omega(\log n)$ lower bound on a set of unsatisfiable integer linear inequalities - the *Simple Pigeonhole Principle* (SPHP) - capturing the core of the counting argument used to prove the PHP efficiently in CP. Since SPHP_n has rank 1 CP proofs, it entails a strong separation between CP rank and SP depth. We then apply the method to PHP_n^m and to $\mathsf{Ts}(K_n,\omega)$.

The *covering method*. The antichain method appears too weak to prove size and depth lower bounds on $\mathsf{Ts}(G,w)$, when G is for example a grid, or a pyramid, or some bounded degree graph. To solve this case, we consider another approach that we call the *covering method*: we reduce the problem of proving that one point in $\mathbb{S}$ survives from all the $\mathsf{slab}(Q)$ in a small proof of $\mathcal{F}$, to the problem that a set of polynomials which *essentially covers* the boolean cube $\{0,1\}^n$ requires at least $\sqrt{n}$ polynomials, which is a well-known problem faced by Alon and Füredi in [1] and by Linial and Radhakrishnan in [62]. For this reduction to work we have to find a high dimensional projection of $\mathbb{S}$ covering the

boolean cube and defined on variables effectively appearing in the proof. We prove that cycles of distance at least 2 in G work properly to this aim on $\mathsf{Ts}(G, \omega)$. Since the grid H_n has many such cycles, we can obtain the lower bound on $\mathsf{Ts}(H_n, \omega)$. The use of Linial and Radhakrishnan's result is not new in proof complexity. Part and Tzameret in [69], independently of us, were using this result in a completely different way from us in the proof system $\mathsf{Res}(\oplus)$ handling clauses over parity equations, and not relying on integer linear inequalities and geometrical reasoning.

The results in Section 2.4 below originally appealed to a lower bound on the size of 'essential covers' given in [62]. Yehuda and Yehudayoff in [88] slightly improved the results of [62] with the consequence, noticed in their paper too, that our size lower bounds for $\mathsf{Ts}(H_n, \omega)$ over a grid graph by what we call the covering method is in fact superlinear in n.

This chapter is organized as follows: We give the preliminary definitions in the next section and then we move to other sections: one on the lower bounds by the antichain method and the other on lower bounds by the covering method. We then give a more specific result for Tseitin Principles, and then finish by pointing out room for improvement.

2.2 Preliminaries

We use $[n]$ for the set $\{1, 2, \ldots, n\}$, $\mathbb{Z}/2$ for $\mathbb{Z} \cup (\mathbb{Z} + \frac{1}{2})$ and $\mathbb{Z}^+$ for $\{1, 2, \ldots\}$.

2.2.1 Restrictions

Let $V = \{x_1, \ldots, x_n\}$ be a set of n variables and let $\mathbf{ax} \leq b$ be a linear integer inequality. We say that a variable x_i *appears in*, or is *mentioned by* a query $Q = (\mathbf{a}, b)$ if $a_i \neq 0$ and *does not appear* otherwise.

A *restriction* ρ is a function $\rho : D \to \{0, 1\}$, $D \subseteq V$. A restriction acts on a half-plane $\mathbf{ax} \leq b$ setting the x_i's according to ρ. Notice that the variables $x_i \in D$ do not appear in the restricted half-plane.

By $\mathcal{T}\restriction_\rho$ we mean to apply the restriction ρ to all the queries in a SP proof $\mathcal{T}$. The tree $\mathcal{T}\restriction_\rho$ defines a new SP proof: if some $Q\restriction_\rho$ reduces to $0 \leq -b$, for some $b \geq 1$, then that node becomes a leaf in $\mathcal{T}\restriction_\rho$. Otherwise in $\mathcal{T}\restriction_\rho$ we simply branch on $Q\restriction_\rho$. Of course the solution space defined by the linear inequalities labelling a path in $\mathcal{T}\restriction_\rho$ is a subset of the

solution space defined by the corresponding path in $\mathcal{T}$. Hence the leaves of $\mathcal{T}\restriction_\rho$ define an infeasible LP.

We work with linear integer inequalities which are a translation of families of CNFs $\mathcal{F}$. Hence when we write $\mathcal{F}\restriction_\rho$ we mean the applications of the restriction ρ to the set of linear integer inequalities defining $\mathcal{F}$.

2.3 The antichain method

This method is based on Sperner's theorem. Using it we can prove depth lower bounds in SP for PHP_n^m and for Tseitin contradictions $\mathsf{Ts}(K_n, \omega)$ over the complete graph. To motivate and explain the main definitions, we use as an example a simplification of the PHP_n^m, the *Simplified Pigeonhole Principle* SPHP_n, which has some interest since (as we will show) it exponentially separates CP rank from SP depth.

2.3.1 Simplified Pigeonhole Principle

As mentioned in Section 2.1.2, the SPHP_n intends to capture the core of the counting argument used to efficiently refute the PHP in CP.

Definition 2.1. The SPHP_n is the following unsatisfiable family of inequalities:

$$\begin{aligned} &\textstyle\sum_{i=1}^{n} x_i \geq 2 \\ &x_i + x_j \leq 1 \qquad \text{for all } i \neq j \in [n] \\ &0 \leq x_i \leq 1 \qquad \text{for all } i \in [n]. \end{aligned}$$

Lemma 2.1. *SPHP_n has a rank 1 CP refutation, for $n \geq 3$.*

Proof. For brevities sake let $S := \sum_{i=1}^{n} x_i$ (so we have $S \geq 2$ as the first axiom). We fix some $i \in [n]$ and sum $x_i + x_j \leq 1$ over all $j \in [n] \setminus \{i\}$ to find

$$\sum_{j \in [n]\setminus i} (x_i + x_j) = (n-2)x_i + (x_i + \sum_{j \in [n]\setminus i} x_j) = (n-2)x_i + S \leq n - 1.$$

We add this to the first axiom $-S \leq -2$ to get

$$x_i \leq \frac{n-3}{n-2}$$

which becomes $x_i \leq 0$ after a single cut. We do this for every i and find $S \leq 0$ - a contradiction when combined with the axiom $S \geq 2$. □

It not hard to see that SPHP_n has depth $O(\log n)$, length $O(n)$ proofs in SP. We are about to give a direct proof, if not just purely for a first informative example of the mechanics of a SP refutation, but please do note that this also follows by appealing to the polynomial size refutations in CP for the PHP_n^m ([21]) and then using Theorem 4.4 in [7] informally stating that 'CP proofs can be balanced in SP'.

Theorem 2.1. *The SPHP_n has a SP refutation of size $O(n)$ and depth $O(\log(n))$.*

Proof. Note that no admissible point for the SPHP_n has any x_i set to 1, as then every other variable is immediately forced to zero and the existential axiom is violated. So then our SP refutation just performs a binary search looking for an x_i set to 1 – if it cannot find such an x_i, we contradict the axiom $\sum_{i=1}^n x_i \geq 2$,

In more detail, the root asks if $\sum_{i=1}^n x_i$ is at least 1 or at most 0. The at most 0 branch directly contradicts the axiom $\sum_{i=1}^n x_i \geq 2$, and so terminates as a leaf (as it is inconsistent as a linear program). The at least 1 branch asks if $\sum_{i=1}^{\lfloor n/2 \rfloor} x_i$ is again at least 1 or at most 0. If this is at most 0, we must have that $\sum_{i=\lfloor n/2 \rfloor+1}^n x_i \geq 1$, and so in either case we have halved the range of the summation containing some x_i hypothetically set to 1. □

We will now prove that this depth lower bound is tight.

2.3.2 Sperner's Theorem

Let $\mathbf{a} \in \mathbb{R}^n$. The *width* $w(\mathbf{a})$ of $\mathbf{a}$ is the number of non-zero coordinates in $\mathbf{a}$. The width of a query $(\mathbf{a}, b)$ is $w(\mathbf{a})$, and the width of a SP refutation is the minimum width of its queries.

Let $n \in \mathbb{N}$. Fix $W \subseteq [0,1] \cap \mathbb{Q}^+$ of finite size $k \geq 2$ and insist that $0 \in W$. The W's we will work with here are $\{0, 1/2\}$ and $\{0, 1/2, 1\}$.

Definition 2.2. A (n, W)-word is an element in W^n.

For two vectors $x, y \in \mathbb{R}^d$, say that $x \leq y$ *in the pointwise ordering* if $x_i \leq y_i$ for all $1 \leq i \leq d$. We consider the following extension of Sperner's theorem.

Theorem 2.2 ([66], Theorem 1.4)**.** *Fix any $t \geq 2, t \in \mathbb{N}$. For all $f \in \mathbb{N}$, with the pointwise ordering of $[t]^f$, any antichain has size at most $t^f \sqrt{\frac{6}{\pi(t^2-1)f}}(1 + o(1))$.*

We will use the simplified bound that any antichain $\mathcal{A}$ has size $|\mathcal{A}| \leq \frac{t^f}{\sqrt{f}}$.

Lemma 2.2. *Let* $\mathbf{a} \in \mathbb{Z}^n$ *and* $|W| = k \geq 2$. *The number of* (n, W)*-words* $\mathbf{s}$ *such that* $\mathbf{as} = b$, *where* $b \in \mathbb{Q}$, *is at most* $\frac{k^n}{\sqrt{w(\mathbf{a})}}$.

Proof. Define $I_{\mathbf{a}} = \{i \in [n] : a_i \neq 0\}$. Let $\preceq$ be the partial order over $W^{I_{\mathbf{a}}}$ where $\mathbf{x} \preceq \mathbf{y}$ if $x_i \leq y_i$ for all i with $a_i > 0$ and $x_i \geq y_i$ for the remaining i with $a_i < 0$. Clearly the set of solutions (restricted to indices in $I_{\mathbf{a}}$) to $\mathbf{as} = b$ forms an antichain under $\preceq$. Noting that $\preceq$ is isomorphic to the typical pointwise ordering on $W^{I_{\mathbf{a}}}$, we appeal to Theorem 2.2 to upper bound the number of solutions in $W^{I_{\mathbf{a}}}$ by $\frac{k^{w(\mathbf{a})}}{\sqrt{w(\mathbf{a})}}$, each of which corresponds to at most $k^{n-w(\mathbf{a})}$ vectors in W^n. □

2.3.3 Large admissibility

A (n, W)-word s is *admissible* for an unsatisfiable set of integer linear inequalities $\mathcal{F}$ over n variables if s satisfies all constraints of $\mathcal{F}$. A set of (n, W)-words is admissible for $\mathcal{F}$ if all its elements are admissible. $\mathcal{A}(\mathcal{F}, W)$ is the set of all admissible (n, W)-words for $\mathcal{F}$.

The interesting sets W for an unsatisfiable set of integer linear inequalities $\mathcal{F}$ are those such that almost all (n, W)-words are admissible for $\mathcal{F}$. We will apply our method on sets of integer linear inequalities which are a translation of unsatisfiable CNF's generated over a given domain. Typically these formulas on a size n domain have a number of variables which is not exactly n but a function of n, $\nu(n) \geq n$. (For example, the PHP_n has $\nu(n) = n^2$ variables.) Hence for the rest of this section we consider $\mathscr{F} := \{\mathcal{F}_n\}_{n\in\mathbb{N}}$ as a family of sets of unsatisfiable integer linear inequalities, where $\mathcal{F}_n$ has $\nu(n) \geq n$ variables. We call $\mathscr{F}$ an *unsatisfiable family*.

Consider then the following definition (recalling that we denote $k = |W|$):

Definition 2.3. $\mathscr{F}$ is *almost full* if $|\mathcal{A}(\mathcal{F}_n, W)| \geq k^{\nu(n)} - o(k^{\nu(n)}) = (1 - o(1))k^{\nu(n)}$, that is, if (asymptotically) almost every point is feasible.

Notice that, because of the o notation, Definition 2.3 might be not necessarily be meaningful for all $n \in \mathbb{N}$, but only starting from some $n_{\mathscr{F}}$.

Definition 2.4. Given some almost full family $\mathscr{F}$ (over $\nu(n)$ variables) we let $n_{\mathscr{F}}$ be the natural number with

$$\frac{k^{\nu(n)}}{|\mathcal{A}(\mathcal{F}_n, W)|} \leq 2 \quad \text{for all} \quad n \geq n_{\mathscr{F}}.$$

As an example we prove SPHP is almost full (notice that in the case of SPHP_n, $\nu(n) = n$).

Lemma 2.3. SPHP_n *is almost full when* $W = \{0, 1/2\}$.

Proof. Let U be the set of all (n, W)-words with at least four coordinates set to $1/2$. U is admissible for SPHP_n since inequalities $x_i + x_j \leq 1$ are always satisfied for any value in W and inequalities $x_1 + \ldots + x_n \geq 2$ are satisfied by all points in U which contain at least four $1/2$s. By a simple counting argument, in U there are at least $2^n - 4n^3 = 2^n - o(2^n)$ admissible (n, W)-words. □

Lemma 2.4. *Let* $\mathscr{F} = \{\mathcal{F}_n\}_{n\in\mathbb{N}}$ *be an almost full unsatisfiable family, where* $\mathcal{F}_n$ *has* $\nu(n)$ *variables. Further let* $\mathcal{T}$ *be a SP refutation of* $\mathcal{F}$ *of width* w. *If* $n \geq n_{\mathscr{F}}$ *then* $|\mathcal{T}| = \Omega(\sqrt{w})$.

Proof. We estimate at what rate the slab of the queries in $\mathcal{T}$ rule out admissible points in U. Let ℓ be the least common multiple of the denominators in W. Every (n, W)-word x falling in the slab of some query $(\mathbf{a}, b)$ satisfies one of ℓ equations $\mathbf{a}x = b + i/\ell, 1 \leq i < \ell$ (as $\mathbf{a}$ is integral). Note that as $|W|$ is a constant independent of n, so is ℓ.

Since all the queries in $\mathcal{T}$ have width at least w, according to Lemma 2.2, each query in $\mathcal{T}$ rules out at most $\ell \cdot \frac{k^{\nu(n)}}{\sqrt{w}}$ admissible points. By Fact 1.2 no point survives at the leaves, in particular the admissible points. Then it must be that

$$|\mathcal{T}|\ell \cdot \frac{k^{\nu(n)}}{\sqrt{w}} \geq |\mathcal{A}(\mathcal{F}_n, W)| \quad \text{which means} \quad |\mathcal{T}|\ell \cdot \frac{k^{\nu(n)}}{|\mathcal{A}(\mathcal{F}_n, W)|} \geq \sqrt{w}$$

We finish by noting that, by the assumption $n \geq n_{\mathscr{F}}$, and then by Definition 2.4, we have $2 \geq \frac{k^{\nu(n)}}{|\mathcal{A}(\mathcal{F}_n, W)|}$, so $|\mathcal{T}| \geq \sqrt{w}/(2\ell) \in \Omega(\sqrt{w})$. □

2.3.4 Main theorem

We focus on restrictions ρ that after applied to an unsatisfiable family $\mathscr{F} = \{\mathcal{F}_n\}_{n\in\mathbb{N}}$, reduce the set $\mathcal{F}$ to another set in the same family.

Definition 2.5. Let $\mathscr{F} = \{\mathcal{F}_n\}_{n\in\mathbb{N}}$ be an unsatisfiable family and c a positive constant. $\mathscr{F}$ is *c-self-reducible* if for any set V of variables, with $|V| = v < n/c$, there is a restriction ρ with domain $V' \supseteq V$, such that $\mathcal{F}_n\restriction_\rho = \mathcal{F}_{n-cv}$ (up to renaming of variables).

Let us motivate the definition with an example.

Lemma 2.5. SPHP_n *is* 1*-self-reducible.*

Proof. Whatever set of variables x_i, $i \in I \subset [n]$ we consider, it is sufficient to set x_i to 0 to fulfil Definition 2.5. □

Theorem 2.3. *Let $\mathscr{F} := \{\mathcal{F}_n\}_{n\in\mathbb{N}}$ be an unsatisfiable set of integer linear inequalities which is almost full and c-self-reducible, for some constant c. If $\mathcal{F}_n$ defines a feasible* LP *whenever $n > n_{\mathscr{F}}$, then for n large enough, the shortest SP proof of $\mathcal{F}_n$ is of length $\Omega(\sqrt[4]{n})$.*

Proof. Take any SP proof $\mathcal{T}$ refuting $\mathcal{F}_n$ and fix $t = \sqrt[4]{n}$.

The proof proceeds by stages $i \geq 0$ where $\mathcal{T}_0 = \mathcal{T}$. The stages will go on while the invariant property (which at stage 0 is true since $n > n_{\mathscr{F}}$ and c a positive constant)

$$n - ict^3 > \max\{n_{\mathscr{F}}, n(1 - 1/c)\}$$

holds.

At the stage i we let $\Sigma_i = \{(\mathbf{a}, b) \in \mathcal{T}_i : w(\mathbf{a}) \leq t^2\}$ and $s_i = |\Sigma_i|$. If $s_i \geq t$ the claim is trivially proven. If $s_i = 0$, then all queries in $\mathcal{T}_i$ have width at least t^2 and by Lemma 2.4 (which can be applied since $n - ict^3 > n_{\mathscr{F}}$) the claim is proven (for n large enough).

So assume that $0 < s_i < t$. Each of the queries in Σ_i involves at most t^2 nonzero coefficients, hence in total they mention at most $s_i t^2 \leq t^3$ variables. Extend this set of variables to some V' in accordance with Definition 2.5 (which can be done since, by the invariant, $ict^3 < n/c$). Set all these variables according to self-reducibility of $\mathcal{F}$ in a restriction ρ_i and define $\mathcal{T}_{i+1} = \mathcal{T}_i \restriction_{\rho_i}$. Note that by Definition 2.5 and by that of restriction, $\mathcal{T}_{i+1}$ is a SP refutation of $\mathcal{F}_{n-ict^3}$ and we can go on with the next stage. (Also note that we do not hit an empty refutation this way, due to the assumption that $\mathcal{F}_n$ defines a feasible LP.)

Assume that the invariant does not hold. If this is because $n - ict^3 < n_{\mathscr{F}}$ then, as each iteration destroys at least one node,

$$|\mathcal{T}| \geq i > \frac{n - n_{\mathscr{F}}}{ct^3} \in \Omega(n^{1/4}).$$

If this is because $n - ict^3 < n - n/c$, then again for the same reason it holds that

$$|\mathcal{T}| \geq i > \frac{n}{c^2 n^{3/4}} \in \Omega(n^{1/4}).$$

□

Using Lemmas 2.3 and 2.5 and the previous Theorem we get:

Corollary 2.1. *The length of any SP refutation of* SPHP_n *is $\Omega(\sqrt[4]{n})$. Hence the minimal depth is $\Omega(\log n)$.*

2.3.5 Lower bounds for the Pigeonhole Principle

We present a lower bound for PHP_n^m closely following that for SPHP_n, in which we largely ignore the diversity of different pigeons (which makes the principle rather like SPHP_n).

In this subsection we fix $W = \{0, 1/2\}$, and for the sake of brevity refer to (n, W)-words as *biwords*.

In this section we fix m to be $n + d$, for any fixed $d \in \mathbb{N}$ at least one.

Lemma 2.6. *The* PHP_n^{n+d} *is almost full (as defined in Definition 2.3).*

Proof. We show that there are at least 2^{mn-1} admissible biwords (for sufficiently large n). For each pigeon i, there are admissible valuations to holes so that, so long as at least two of these are set to $1/2$, the others may be set to anything in $\{0, 1/2\}$. This gives at least $2^n - (n+1)$ possibilities. Since the pigeons are independent, we obtain at least $(2^n - (n+1))^m$ biwords. Now this is $2^{mn}\left(1 - \frac{n+1}{2^n}\right)^m$ where $\left(1 - \frac{n+1}{2^n}\right)^m \sim e^{\frac{-(n+1)m}{2^n}}$ whence, $\left(1 - \frac{n+1}{2^n}\right)^m \geq e^{\frac{-(n+2)m}{2^n}}$ for sufficiently large n. It follows there is a constant c so that:

$$2^{mn}\left(1 - \frac{n+1}{2^n}\right)^m \geq 2^{mn - \frac{c(n+2)m}{2^n}} \geq 2^{mn-1}$$

for sufficiently large n. □

Lemma 2.7. *The* PHP_n^{n+d} *is* 1*-self-reducible.*

Proof. We are given some set I of variables from PHP_n^{n+d}. These variables will mention some set of holes $H := \{j : P_{i,j} \in I \text{ for some i}\}$ and similarly a set of pigeons P. Each of P, H have size at most $|I|$ and we extend them both arbitrarily to have size exactly $|I|$. Our restriction matches P and H in any way and then sets any other variable mentioning a pigeon in P or a hole in H to 0. □

Theorem 2.4. *The length of any SP refutation of* PHP_n^{n+d} *is* $\Omega(n^{1/4})$.

Proof. Note that the all $1/2$ point is feasible for PHP_n^{n+d}. Then with Lemma 2.6 and Lemma 2.7 in hand we meet all the prerequisites for Theorem 2.3.

□

By simply noting that a SP refutation is a binary tree, we get the following corollary.

Corollary 2.2. *The SP depth of the* PHP_n^{n+d} *is* $\Omega(\log n)$.

2.3.6 Lower bounds for Tseitin contradictions over the complete graph

Definition 2.6. Recall from section 1.1.4 that, for a graph $G = (V, E)$ along with a charging function $\omega : V \to \{0, 1\}$ satisfying $\sum_{v \in V} \omega(v) = 1 \mod 2$, the *Tseitin contradiction* $\mathsf{Ts}(G, \omega)$ is the set of linear inequalities which translate the CNF encoding of

$$\sum_{\substack{e \in E \\ e \ni v}} x_e = \omega(v) \mod 2. \tag{2.1}$$

for every $v \in V$, where the variables x_e range over the edges $e \in E$.

The following fact comes from the observation that, when expanding Equation (2.1) into CNF, every clause mentions every variable:

Fact 2.1. Let e, e' be incident to a vertex v in G, and ω an odd charging function. For any nonnegative assignment to the variables of $\mathsf{Ts}(G, \omega)$ setting $x_e, x_{e'}$ to $1/2$, the parity constraint at v is satisfied.

In this subsection we consider $\mathsf{Ts}(K_n, \omega)$ and ω will always be an odd charging for K_n. We let $N := \binom{n}{2}$ and we fix $W = \{0, 1/2, 1\}$, $k = 3$ and for the sake of brevity refer to (n, W)-words as *triwords*. We will abuse slightly the notation of Section 2.3.3 and consider the family $\{\mathsf{Ts}(K_n, \omega)\}_{n \in \mathbb{N},\, \omega \text{ odd}}$ as a single parameter family in n. The reason we can do this is because the following proofs of almost fullness and self reducibility do not depend on ω at all (so long as it is odd, which we will always ensure).

Lemma 2.8. $\mathsf{Ts}(K_n, \omega)$ *is almost full.*

Proof. We show that $\mathsf{Ts}(K_n, \omega)$ has at least $c3^N$ admissible triwords, for any constant $0 < c < 1$ and n large enough. We define the assignment ρ setting all edges (i.e. x_e) to a value in $W = \{0, 1, 1/2\}$ independently and uniformly at random, and inspecting the probability that some fixed constraint for a node v is violated by ρ.

Clearly if at least 2 edges incident to v are set to $1/2$ its constraint is satisfied. If none of its incident edges are set to $1/2$ then it is satisfied with probability $1/2$. Let $A(v)$ be the event "*no edge incident to* v *is set to* $1/2$ *by* ρ" and let $B(v)$ be the event that "*exactly one edge incident to* v *is set to* $1/2$ *by* ρ". Then:

$$\Pr[v \text{ is violated}] \leq \frac{1}{2}\Pr[A(v)] + \Pr[B(v)] = \frac{1}{2}\frac{2^{n-1}}{3^{n-1}} + \frac{(n-1)2^{n-2}}{3^{n-1}} = n\frac{2^{n-2}}{3^{n-1}}.$$

Therefore, by a union bound, the probability that there exists a node with violated parity is bounded above by $n^2 \frac{2^{n-2}}{3^{n-1}}$, which approaches 0 as n goes to infinity. □

Lemma 2.9. $\mathsf{Ts}(K_n, \omega)$ *is* 2*-self-reducible.*

Proof. We are given some set of variables I. Each variable mentions 2 nodes, so extend these mentioned nodes arbitrarily to a set S of size exactly $2|I|$, which we then hit with the following restriction: if S is evenly charged, pick any matching on the set $\{s \in S : w(s) = 1\}$, set those edges to 1, and set any other edges involving some vertex in S to 0. Otherwise (if S is oddly charged) pick any $l \in \{s \in S : w(s) = 1\}$ and $r \in [n] \setminus S$ and set x_{lr} to 1. $\{s \in S : w(s) = 1\} \setminus l$ is now even so we can pick a matching as before. And as before we set all other edges involving some vertex in S to 0. In the first case the graph induced by $[n] \setminus S$ must be oddly charged (as the original graph was). In the second case this induced graph was originally evenly charged, but we changed this when we set x_{lr} to 1. □

Lemma 2.10. *For any oddly charged* ω *and* n *large enough, all SP refutations of* $\mathsf{Ts}(K_n, \omega)$ *have length* $\Omega(\sqrt[4]{n})$.

Proof. We have that the all 1/2 point is feasible for $\mathsf{Ts}(K_n, \omega)$. Then we can simply apply Theorem 2.3. □

Corollary 2.3. *The depth of any SP refutation of* $\mathsf{Ts}(K_n, \omega)$ *is* $\Omega(\log n)$.

2.3.7 Lower bound for the Least Ordering Principle

Lemma 2.11. *For any* $X \subseteq [n]$ *of size at most* $n - 3$*, there is an admissible point for* LOP_n *integer on any edge mentioning an element in* X.

Proof. Let $\preceq$ be any total order on the elements in X. Our admissible point x will be

$$x(P_{i,j}) = \begin{cases} 1 & \text{if } i,j \in X \text{ and } i \preceq j, \text{ or if } i \notin X, j \in X \\ 0 & \text{if } i,j \in X \text{ and } j \preceq i, \text{ or if } i \in X, j \notin X \\ 1/2 & \text{otherwise (if } i,j \notin X). \end{cases}$$

The existential axioms $\sum_{i=1, i \neq j}^{n} P_{i,j}$ are always satisfied - if $j \in X$ then there is some $i \notin X$ with $P_{i,j} = 1$, and otherwise there are at least two distinct $i, k \neq j \in X$ with $P_{i,j}, P_{k,j} = 1/2$. For the transitivity axioms $P_{i,k} - P_{i,j} - P_{j,k} \geq 1$, note that if 2 or more of i, j, k are not in X there are at least 2 variables set to 1/2, and otherwise it is set in a binary fashion to something consistent with a total order.

□

We will assume that a SP refutation $\mathcal{T}$ of LOP_n only involves variables $P_{i,j}$ where $i < j$ - this is without loss of generality as we can safely set $P_{j,i}$ to $1 - P_{i,j}$ whenever $i > j$, and will often write $P_{\{i,j\}}$ for such a variable. We consider the underlying graph of the support of a query, i.e. an undirected graph with edges $\{i, j\}$ for every variable $P_{\{i,j\}}$ that appears with non-zero coefficient in the query.

For some function $f(n)$, we say the query is $f(n)$-*wide* if the smallest edge cover of its graph has at least $f(n)$ nodes. A query that is not $f(n)$-wide is $f(n)$-*narrow*. The next lemma works much the same as Theorem 2.3.

Lemma 2.12. *Fix $\epsilon > 0$ and suppose we have some SP refutation $\mathcal{T}$ of LOP_n, where $|\mathcal{T}| \leq n^{\frac{1-\epsilon}{4}}$. Then, if n is large enough, we can find some SP refutation $\mathcal{T}'$ of $\mathsf{LOP}_{c \cdot n}$, where c is a positive universal constant that may be taken arbitrarily close to 1, $\mathcal{T}'$ contains only $n^{3/4}$-wide queries, and $|T'| \leq |\mathcal{T}|$.*

Proof. We iteratively build up an initially empty restriction ρ. At every stage ρ imposes a total order on some subset $X \subseteq [n]$ and places the elements in X above the elements not in X. So ρ sets every edge not contained entirely in $[n] \setminus X$ to something binary, and $\mathsf{LOP}_n\restriction_\rho = \mathsf{LOP}_{n-|X|}$ (up to a renaming of variables).

While there exists a $n^{3/4}$-narrow query $q \in \mathcal{T}\restriction_\rho$ we simply take its smallest edge cover, which has size at most $n^{3/4}$ by definition, and add its nodes in any fashion to the total order in ρ. Now all of the variables mentioned by $q \in \mathcal{T}\restriction_\rho$ are fully evaluated and q is redundant. We repeat this at most $n^{\frac{1-\epsilon}{4}}$ times (as $|\mathcal{T}| \leq n^{\frac{1-\epsilon}{4}}$ and each iteration renders at least one query in $\mathcal{T}$ redundant). At each stage we grow the domain of the restriction by at most $n^{3/4}$, so the domain of ρ is always bounded by $n^{1-\epsilon/4}$. We also cannot exhaust the tree $\mathcal{T}$ in this way, as otherwise $\mathcal{T}$ mentioned at most $n^{1-\epsilon/4} < n-3$ elements and by Lemma 2.11 there is an admissible point not falling in any slab of $\mathcal{T}$, violating Lemma 1.2.

When this process finishes we are left with a $n^{3/4}$-wide refutation $\mathcal{T}'$ of $\mathsf{LOP}_{n-n^{1-\epsilon/4}}$. As ϵ was fixed we find that as n goes to infinity $n - n^{1-\epsilon/4}$ tends to n. □

Lemma 2.13. *Let $d \leq (n-3)/2$. Given any disjoint set of pairs $D = \{\{l_1, r_1\}, \ldots, \{l_d, r_d\}\}$ (where without loss of generality $l_i < r_i$ in $[n]$ as natural numbers) and any binary assignment $b \in \{0,1\}^D$, the assignment x_b with*

$$x_b(P_{\{i,j\}}) = \begin{cases} b(\{l_k, r_k\}) & \text{if } \{i,j\} = \{l_k, r_k\} \in X \text{ for some } k \\ 1/2 & \text{otherwise} \end{cases}$$

is admissible.

Proof. The existential axioms $\sum_{i=1,i\neq j}^{n} P_{i,j}$ are always satisfied, as for any j there are at least $n-2$ $i \in [n]$ different from j with $P_{i,j} = 1/2$. For the transitivity axioms $P_{i,k} - P_{i,j} - P_{j,k} \geq 1$, note that due to the disjointness of D at least two variables on the left hand side are set to $1/2$. □

Theorem 2.5. *Fix some $\epsilon > 0$ and let $\mathcal{T}$ any SP refutation of LOP_n. Then, for n large enough, $|\mathcal{T}| \in \Omega(n^{\frac{1-\epsilon}{4}})$.*

Proof. Suppose otherwise - then, by Lemma 2.12, we can find some $\mathcal{T}'$ refuting LOP_{cn}, with $|\mathcal{T}'| \leq |\mathcal{T}|$, every query $n^{3/4}$-wide, and c independent of n. We greedily create a set of pairs D by processing the queries in $\mathcal{T}'$ one by one and choosing in each a matching of size $n^{1/2}$ disjoint from the elements appearing in D - this always succeeds, as at every stage $|D| \in O(n^{\frac{1-\epsilon}{4}} \cdot n^{1/2})$ and involves at most $O(2n^{\frac{3-\epsilon}{4}}) < n^{3/4} - n^{1/2}$ elements.

So by Lemma 2.13, after setting every edge not in D to $1/2$, we have some set of linear polynomials $\mathcal{R} = \{a(x) = \mathbf{a}x - b - 1/2 : (\overline{a}, b) \in \mathcal{T}'\}$ covering the hypercube $\{0,1\}^D$, where every polynomial $p \in \mathcal{R}$ mentions at least $n^{1/2}$ edges. By Lemma 2.2 each such polynomial in $\mathcal{R}$ rules out at most $2^{|D|}/n^{1/4}$ points, and so we must have $|\mathcal{T}| \geq |T'| \geq |\mathcal{R}| \geq n^{1/4}$. □

2.4 The covering method

Definition 2.7. A set L of linear polynomials with real coefficients is said to be a *cover* of the cube $\{0,1\}^n$ if for each $v \in \{0,1\}^n$, there is a $p \in L$ such that $p(v) = 0$.

In [62] Linial and Radhakrishnan considered the problem of the minimal number of hyperplanes needed to cover the cube $\{0,1\}^n$. Clearly every such cube can be covered by the zero polynomial, so to make the problem more meaningful they defined the notion of an *essential covering* of $\{0,1\}^n$.

Definition 2.8 ([62])**.** A set L of linear polynomials with real coefficients is said to be an *essential cover* of the cube $\{0,1\}^n$ if

(E1) L is a cover of $\{0,1\}^n$,

(E2) no proper subset of L satisfies (E1), that is, for every $p \in L$, there is a $v \in \{0,1\}^n$ such that p alone takes the value 0 on v, and

(E3) every variable appears (in some monomial with non-zero coefficient) in some polynomial of L.

They then proved that any essential cover E of the hypercube $\{0,1\}^n$ must satisfy $|E| \geq \sqrt{n}$. We will use the slightly strengthened lower bound given in [88]:

Theorem 2.6. *Any essential cover L of the cube with n coordinates satisfies $|L| \in \Omega(n^{0.52})$.*

We will need an auxiliary definition and lemma.

Definition 2.9. Let L be a cover of $\{0,1\}^I$ for some index set I. Some subset L' of L is an *essentialisation* of L if L' also covers $\{0,1\}^I$ but no proper subset of it does.

Lemma 2.14. *Let L be a cover of the cube $\{0,1\}^n$ and L' be any essentialisation of L. Let M' be the set of variables appearing with nonzero coefficient in L'. Then L' is an essential cover of $\{0,1\}^{M'}$.*

Proof.

(E1) Given any point $x \in \{0,1\}^{M'}$, we can extend it arbitrarily to a point $x' \in \{0,1\}^M$. Then there is some $p \in L'$ with $p(x') = 0$ - but $p(x') = p(x)$, as p doesn't mention any variable outside of M'.

(E2) Similarly to the previous point, this will follow from the fact that if some set $\mathcal{T}$ covers a hypercube $\{0,1\}^I$, it also covers $\{0,1\}^{I'}$ for any $I' \supseteq I$.

Suppose some proper subset $L'' \subset L'$ covers $\{0,1\}^{M'}$, then it covers $\{0,1\}^M$ - but we picked L' to be a minimal set with this property.

(E3) We defined M' to be the set of variables appearing with nonzero coefficient in L'.

□

2.4.1 The covering method and Tseitin

Let H_n denote the $n \times n$ grid graph. Fix some ω with odd charge and a SP refutation $\mathcal{T}$ of $\mathsf{Ts}(H_n, \omega)$. Lemma 1.2 tells us that for every point x admissible for $\mathsf{Ts}(H_n, \omega)$, there exists a query $(\mathbf{a}, b) \in \mathcal{T}$ such that $b < \mathbf{a}x < b + 1$. In this section we will only consider admissible points with entries in $\{0, 1/2, 1\}$, turning the slab of a query $(\mathbf{a}, b)$ into the solution set of the single linear equation $\mathbf{a} \cdot x = b + 1/2$. So we consider $\mathcal{T}$ as a set of such equations.

We say that an edge of H_n is *mentioned* in $\mathcal{T}$ if the variable x_e appears with non-zero coefficient in some query in $\mathcal{T}$. We can see H_n as a set of $(n-1)^2$ squares (4-cycles), and we can index them as if they were a Cartesian grid, starting from 1. Let S be the set of $\lfloor (n/3)^2 \rfloor$ squares in H_n gotten by picking squares with indices that become 2 (mod 3). This ensures that every two squares in S in the same row or column have at least two other squares between them, and that no selected square is on the perimeter.

We will assume without loss of generality that n is a multiple of 3, so $|S| = (n/3)^2$. Let $K = \bigcup_{t \in S} t$ be the set of edges mentioned by S, and for some $s \in S$, let $K_s := \bigcup_{t \in S, t \neq s} t$ be the set of edges mentioned in S by squares other than s.

Lemma 2.15. *For every $s \in S$ we can find an admissible point $b_s \in \{0, 1/2, 1\}^{E(H_n)}$ such that*

1. *$b_s(x_e) = 0$ for all $e \in K_s$, and*
2. *b_s is fractional only on the edges in s.*

Proof. We state a specialization of Corollary 1.1.

Fact 2.2. For each vertex v in H_n there is a totally binary assignment, called v-critical in [85], satisfying all parity axioms in $\mathsf{Ts}(H_n, \omega)$ except the parity axiom of node v.

Pick any corner c of s. Let b_s be the result of taking any c-critical assignment of the variables of $\mathsf{Ts}(H_n, \omega)$ and setting the edges in s to $1/2$. b_s is admissible, as c is now adjacent to two variables set to $1/2$ (so its originally falsified parity axiom becomes satisfied) and every other vertex is either unaffected or also adjacent to two $1/2$s. While b_s sets some edge $e \in K_s$ to 1, flip all of the edges in the unique other square containing e. This other square always exists (as no square touches the perimeter) and also contains no other edge in K_s (as there are at least two squares between any two squares in S). Flipping the edges in a cycle preserves admissibility, as every vertex is adjacent to 0 or 2 flipped edges. □

Definition 2.10. Let $V_S := \{v_s : s \in S\}$ be a set of new variables. For $s \in S$ define the substitution h_s, taking the variables of $\mathsf{Ts}(H_n, \omega)$ to $V_S \cup \{0, 1/2, 1\}$, as

$$h_s(x_e) := \begin{cases} b_s(e) & \text{if } e \text{ is not mentioned in } S, \text{ or if } e \text{ is mentioned by } s, \\ v_t & \text{if } e \text{ is mentioned by some square } t \neq s \in S. \end{cases}$$

(where b_s is from Lemma 2.15).

Definition 2.11. Say that a linear polynomial $p = c + \sum_{e \in E(H_n)} \mu_e x_e$ with coefficients $\mu_e \in \mathbb{Z}$ and some constant part $c \in \mathbb{R}$ has *odd coefficient in* $X \subseteq E(H_n)$ if $\sum_{e \in X} \mu_e$ is an odd integer, and otherwise, we say it has *even coefficient*. Given some polynomial p in the variables x_e of Tseitin, and some square $s \in S$, let p_s be the polynomial in variables V_S gotten by applying the substitution $x_e \to h_s(x_e)$. Also, for any set of polynomials $\mathcal{T}$ in the variables x_e let $\mathcal{T}_s := \{p_s : p \in \mathcal{T}, p \text{ has odd coefficient in } s\}$.

Given some assignment $\alpha \in \{0,1\}^{V_S \setminus \{v_s\}}$, and some h_s as in Definition 2.10, we let $\alpha(h_s)$ be the assignment to the variables of $\mathsf{Ts}(H_n, \omega)$ gotten by replacing the v_t in the definition of h_s by $\alpha(v_t)$.

Lemma 2.16. *Let $s \in S$. For all $2^{|S|-1}$ settings α of the variables in $V_S \setminus \{s\}$, $\alpha(h_s)$ is admissible.*

Proof. When $\alpha(v_t)$ is all 0, $h_s = b_s$ is admissible (by Lemma 2.15). Toggling some v_t only has the effect of flipping every edge in a cycle, which preserves admissibility. □

Lemma 2.17. *$\mathcal{T}_s$ covers $\{0,1\}^{V_S \setminus \{s\}}$.*

Proof. For every setting of $\alpha \in \{0,1\}^{V_S \setminus \{s\}}$, $\alpha(h_s)$ as defined above is admissible and therefore covered by some $p \in \mathcal{T}$, which has constant part $1/2 + b$ for some $b \in \mathbb{Z}$. Furthermore, as $\alpha(h_s)$ sets every edge in s to $1/2$, every such p must have odd coefficient in front of s - otherwise

$$p(\alpha(h_s)) = 1/2 + b + (1/2)\left(\sum_{e \in s} \mu_e\right) + \sum_{e \notin s} \mu_e \alpha(h_s)(x_e)$$

can never be zero, as the $1/2$ is the only non integral term in the summation. □

Theorem 2.7. *Any SP refutation $\mathcal{T}$ of $\mathsf{Ts}(H_n, \omega)$ must have $|\mathcal{T}| \in \Omega(n^{1.04})$.*

Proof. We are going to find a set of pairs $(L_1, M_1), (L_2, M_2), \ldots, (L_q, M_q)$, where the L_i are pairwise disjoint nonempty subsets of $\mathcal{T}$, the M_i are subsets of V_S, and for every i there is some $s_i \in S \setminus \bigcup_{j=1}^{q} M_j$ such that $|(L_i)_{s_i}| \geq |M_i|^{0.52}$. These pairs will also satisfy the property that

$$\{s_i : 1 \leq i \leq q\} \cup \bigcup_{i=1}^{q} M_i = S. \tag{2.2}$$

As $|S| = (n/3)^2$ this would imply that $\sum_{i=1}^{q} |M_i| \geq (n/3)^2 - q$. If $q \geq (n/3)^2/2$, then (as the L_i are nonempty and pairwise disjoint) we have $|\mathcal{T}| \geq (n/3)^2/2 \in \Omega(n^{1.04})$. Otherwise $\sum_{i=1}^{q} |M_i| \geq (n/3)^2/2$, and as (by Theorem 2.6, and because $|L_i| \geq |(L_i)_{s_i}|$) each $|L_i| \geq |M_i|^{0.52}$,

$$|\mathcal{T}| \geq \sum_{i=1}^{q} |L_i| \geq \sum_{i=1}^{q} |M_i|^{0.52} \geq \left(\sum_{i=1}^{q} |M_i| \right)^{0.52} \geq \left((n/3)^2/2 \right)^{0.52} \in \Omega(n^{1.04}). \tag{2.3}$$

We create the pairs by stages. Let $S_1 = S$ and start by picking any $s_1 \in S_1$. By Lemma 2.17 $\mathcal{T}_{s_1}$ covers $\{0,1\}^{V_{S_1} \setminus \{s_1\}}$ and has as an essentialisation E, which will be an essential cover of $\{0,1\}^{V'}$ for some $V' \subseteq V_{S_1} \setminus \{s_1\}$. We create the pair $(L_1, M_1) = (\{p : p_{s_1} \in E\}, V')$ and update $S_2 = S_1 \setminus (V' \cup \{s_1\})$. (Note that V' could possibly be empty - for example, if the polynomial $x_e = 1/2$ appears in $\mathcal{T}$, where $e \in s_1$. In this case however we still have $|L_1| \geq |M_1|^{0.52}$. If V' is not empty we have the same bound due to Theorem 2.6.) If S_2 is nonempty we repeat with any $s_2 \in S_2$, and so on.

We now show that as promised the left hand sides of these pairs partition a subset of $\mathcal{T}$, which will give us the first inequality in Equation (2.3). Every polynomial p with $p_{s_i} \in L_i$ has every v_t mentioned by p_{s_i} removed from S_j for all $j \geq i$, so the only way p could reappear in some later L_j is if $p_{s_j} \in \mathcal{T}_{s_j}$, where v_{s_j} does *not* appear in p_{s_i}. Let $\mu_e, e \in s_j$ be the coefficients of p in front of the four edges of s_j. The coefficient in front of v_{s_j} in p_{s_i} is just $\sum_{e \in s_j} \mu_e$. As v_{s_j} failed to appear this sum is 0 and p does not have the odd coefficient sum it would need to appear in $\mathcal{T}_{s_j}$. □

Corollary 2.4. *Any SP refutation of* $\mathsf{Ts}(H_n, \omega)$ *requires depth* $\Omega(\log n)$ *to refute in Stabbing Planes.*

2.5 Tseitin Principles and circuit rank

Here we describe another method of showing lower bounds for the size of SP refutations of Tseitin Principles. It is simpler than the method of essential coverings just described, and in many cases (such as for the grid) it is stronger - however, it is entirely specific to Tseitin Principles, and shows no obvious potential to generalise.

Definition 2.12. Let G be a graph. A subgraph $E \subset E(G)$ is *Eulerian* if every node in G is adjacent to an even number of edges in E.

Fact 2.3. Let G be a graph. Given two Eulerian subgraphs $E_1, E_2 \subseteq E(G)$, not necessarily induced, their symmetric difference $E_1 \triangle E_2$ is also Eulerian.

Proof. In general the symmetric difference of two sets of even cardinality, in this case two sets of edges adjacent to some node, remains even. □

Definition 2.13. A *cycle basis* B of a graph G is a minimal set of simple cycles such that every Eulerian subgraph of G can be expressed as a symmetric difference of elements in B. The size of any cycle basis, called the *circuit rank* of G, is equal to $|E(G)|-|V(G)|+c$, where c is the number of connected components in G.

A comprehensive survey on cycle bases is [52].
The cycle basis is literally a basis of a vector space (the so-called cycle space, with operation $\triangle$ over the two element field) and so any product $\pi = b_1 \triangle b_2 \triangle \ldots \triangle b_k$ of distinct elements from B with $k \geq 1$ is nonempty. It follows from the definition of symmetric difference that every node in G has even degree in π. So, as before (i.e., as in Section 2.3.6 and Lemma 2.15), by picking a critical assignment failing only the parity clause for a vertex v adjacent to π, we get an admissible point b_π, as every vertex in G is adjacent to an even number of edges in π. In particular, every vertex is adjacent to either 0 such edges, in which case it is not v and has its parity constraint satisfied by b_π being v-critical originally, or is adjacent to at least two edges that are set to $1/2$. (We actually have many candidate b_π, as there are many consistent settings of the integer part of b_π, but we only need one, and any will work for the following proof.)

Theorem 2.8. *Any SP refutation $\mathcal{T}$ of $\mathsf{Ts}(G, \omega)$ must have size at least the circuit rank of G, for any odd charging ω.*

Proof. The proof proceeds in stages, where at the ith stage, we have maintained the invariant that some set of polynomials $\mathcal{T}_i$ covers the set of admissible points $\{b_\pi \ : \ \pi \in B_i\}$ for some set of Eulerian cycles B_i. Initially we set $\mathcal{T}_0$ to be the set of linear equalities associated with the slabs in $\mathcal{T}$ just as before, that is,

$$\mathcal{T}_0 = \{\mathbf{q}(x) = r + 1/2 \ : \ (\mathbf{q}, r) \in \mathcal{T}\}$$

and we let B_0 be any cycle basis of G. As the b_π are triwords, the invariant starts off true (because again as before, if a triword b_π falls in the interval $r < \mathbf{q} \cdot b_\pi < r + 1$, we must have $\mathbf{q} \cdot b_\pi = r + 1/2$).

Suppose we are at the ith stage. We pick any Eulerian cycle $\pi \in B_i$. By the invariant, some polynomial $p \in \mathcal{T}_i$ must kill the point b_π, and in therefore this p must have odd total coefficient in front of the edges in π. (For the sake of clarity, we remind the reader that the b_π are indexed by the original variables of Tseitin, which is to say, edges not cycles. The polynomial p is similarly considered as a linear polynomial in variables corresponding to edges.) Let $B' := B_i \setminus \{\pi\}$ be the remaining elements of B_i and let $S_p \subseteq B'$ be the other elements μ of the cycle basis such that p has odd total coefficient in front of μ. Note:

- if p has odd coefficient in front of some cycle γ, it has even coefficient in front of $\pi \triangle \gamma$, as p had odd coefficient in front of π,
- if p has even coefficient in front of two cycles γ and γ', it has even coefficient in front of $\gamma \triangle \gamma'$, and
- if p has even coefficient in front of some cycle γ, it cannot evaluate to anything noninteger on an admissible point gotten by setting exactly the edges in γ to $1/2$.

So we 'co-sacrifice' π and p by using the single cycle π to render all the other cycles odd for p even instead. We set $B_{i+1} = (B' \setminus S_p) \cup \{\pi \triangle \gamma : \gamma \in S_p\}$, which has size $|B_i| - 1$, as $\pi \triangle \gamma = \pi \triangle \gamma'$ if and only if $\gamma = \gamma'$. Every element in B_{i+1}, and therefore all of their products, has even coefficient in front of p, so p can never be involved in destroying admissible points noninteger only on these subgraphs. Furthermore, a nonempty product can never become the empty graph (as we started with a cycle basis). Then the set $\mathcal{T}_{i+1} := \mathcal{T}_i \setminus \{p\}$, with cardinality exactly $|\mathcal{T}_i| - i$, must cover B_{i+1}, and therefore cannot become empty until B_{i+1} becomes empty. We finish by noting that this only happens when the stage i becomes the circuit rank $|B_0|$. □

2.6 Conclusions and acknowledgements

All of the methods here have a glaring weakness - they *all ignore the entire structure of the SP tree* and flatten it into a disordered set of linear equations. Hence, they can never produce more than a polynomial length lower bound (as the supposed refutation might have a slab $0 < x < 1$ for each variable x in the principle). Given this, it is surprising to the author that the lower bounds produced are ever tight at all, and we believe that incorporating the recursive structure in any way is bound to lead to much stronger length and depth lower bounds. We speculate that one method of correcting this weakness would be to adapt the Res(Lin) prover-delayer game from [69], already working with essential covers, to work for Stabbing Planes. However, we do not attempt this in the present thesis.

We would like to thank Noah Flemming for answering some questions on his paper [7], sending us his manuscript [35], and for comments on a preliminary version of this chapter as it was being prepared for publication.

Chapter 3

Lower bounds for some First Order theories

Let $\mathcal{T}$ be a First Order principle. In the manner described in Section 1.1.1, for every $n \in \mathbb{N}$ we can generate uniformly a propositional CNF $\mathcal{T}_n$ which is satisfiable if and only if $\mathcal{T}$ has a model of size n. If $\mathcal{T}$ has no finite models, which for us will always be the case, we generate propositional contradictions, and can now investigate how the model theory of $\mathcal{T}$ affects the propositional proof complexity of $\mathcal{T}_n$.

A very archetypal result along these lines is the following 'gap theorem', which comes in two parts, a lower bound and an upper bound.

Theorem 3.1 ([28], Theorem 1.2)**.** *Let $\mathcal{T}$ be a first-order principle admitting no finite models. Then, if $\mathcal{T}$ has infinite models, the SA rank of $\mathcal{T}_n$ is $\Omega(n^\epsilon)$, for some $\epsilon > 0$ independent of n. Otherwise, if $\mathcal{T}$ has no models at all, $\mathcal{T}_n$ has constant SA rank.*

In the first section of this chapter, we use machinery from the previous chapter to prove an analogy of this lower bound for SA in SP.

Then, in the second section, we turn to the more powerful Sum Of Squares proof system, and, while we do not have the same success, we do establish some general framework - in particular, we provide some 'canonical' pseudodistribution and find some equivalent conditions for its positive semidefiniteness.

3.1 Stabbing Planes

In this Section we generalise the SP lower bounds for the LOP (Theorem 2.5) and PHP (Theorem 2.4) given in the previous chapter. The generalisation comes from viewing these two principles as just two of many principles that are generated from some FO sentence $\mathcal{T}$ as mentioned in the abstract of this chapter.

In the following Section 3.1.1, we begin by undergoing a number of steps to transform the principle $\mathcal{T}$ into one vulnerable to the antichain framework established in the previous chapter. We call this a 'sanitisation' of the theory. We use this sort of language because, as we will show, any obstacle to the application of the antichain method must come from a type of redundancy or wastefulness in the principle, which we straightforwardly remove. For example, in Lemma 3.1, we show that if the situation in some small finite part of a model fixes some relation to be constant outside of that finite part, then that relation only ever needed to be defined with respect to that small finite part (which we name as constants), and not generically. This allows us to manipulate variables independently and nets us a hypercube.

As we transform the principle $\mathcal{T}$ into some principle $\mathcal{T}'$ by sanitisation, we also transform any refutation D of $\mathcal{T}_n$ into a refutation D' of ${\mathcal{T}'}_{cn}$, with $|D'| \leq |D|$ and $c > 0$ a constant independent of n. After sanitisation $\mathcal{T}$ will almost full as defined in Definition 2.3, as we now find in its solution spaces large hypercubes. Then, in Section 3.1.2, we lower bound the size of $|D'|$, which is now almost susceptible to attack by the antichain method from the previous chapter, but not quite.

To wit, the remaining problem is this: in for example Lemma 2.12, there was a straightforward restriction for the LOP that showed self-reducibility - we just placed the elements we wished to eliminate above all the other elements, rendering any variable mentioning those elements constantly integral, and effectively removing them from D. But for generic $\mathcal{T}$ it is not obviously the case that $\mathcal{T}_n$ is self-reducible in the same way. This was a crucial part of the antichain method.

We work around this in Lemma 3.8 by running a 'bounded re-sanitisation' - instead of immediately making redundant a query with low covering number, we name that cover as a set of constants and re-establish disjointness (which we will note has bounded cost). This may not already destroy the query as it did for the PHP and LOP in the previous chapter, but it does reduce the effective arity of every single variable in that query (as they now must all touch a constant), and for a given query this can only happen a constant number of times (bounded by the arity of the principle).

A note on the aesthetics of the proof. $\mathcal{T}'$ is gotten from $\mathcal{T}$ by addition of some finite number of axioms and a finite extension of the language of $\mathcal{T}$. However all of our axioms are very simple and only ever name constants or set single literals, and the extensions are a similarly simple replacement of literals (see Fact 3.1 and the discussion following Lemma 2.11 for a specific case). In the 'propositional' world of $\mathcal{T}_n$ this is just selecting a slice of the polytope in which we can locate a useful hypercube. An equivalent, more verbose, but perhaps more simplistic proof could show the same lower bound from this angle.

3.1.1 Sanitising the theory

In this section, unemboldened arguments (like the τ_1 in $R_1(\tau_1)$) are to be read as fixed, and can be thought of as the 'concrete arguments' in the literals of the CNF produced by the procedure described in Section 1.1.1, instantiated with elements from $\mathbb{N}$, unless a quantifier claims otherwise. Emboldened parameters are to be read as free variables.

Definition 3.1. Let $\mathcal{L}$ be some vocabulary and let $\mathcal{R} = \{R_1(\tau_1), \ldots, R_k(\tau_k)\}$ be some set of relations from $\mathcal{L}$ instantiated with parameters in $[n]$, for some $n \in \mathbb{N}$. Let $\mathfrak{M}$ be an $\mathcal{L}$-structure and $\rho \in \{0,1\}^{\mathcal{R}}$ be an assignment to $\mathcal{R}$. An injection $\iota : [n] \to \mathfrak{M}$ is *consistent* with ρ if the interpretation by $\mathfrak{M}$ of $R_i(\iota(\tau_i))$ agrees with $\rho(R_i(\tau_i))$ for all $1 \leq i \leq k$ (interpreting 1 as true).

In the previous chapter, it was useful (for example in Lemma 2.13) that variables with disjoint subscripts could be set independently. This is not obviously true for all principles, but we will show that it is 'effectively true infinitely often' in some subset of models, which will give us a set of admissible points containing actionable hypercubes.

Definition 3.2. Let $\mathcal{T}$ be a theory with some finite set of constants $C \subset \mathbb{N}$. A set of relations $\mathcal{R} = \{R_i(\tau_i) : 1 \leq i \leq k\}$ instantiated with parameters in $\mathbb{N}$ is said to be *disjoint* if (a) every tuple τ_i mentions nothing in C and (b) the tuples τ_i, τ_j are pairwise disjoint. A theory $\mathcal{T}$ is said to be *disjoint* if for any disjoint set of relations $\mathcal{R} = \{R_i(\tau_i) : 1 \leq i \leq k\}$ and any assignment $b \in \{0,1\}^{\mathcal{R}}$, there is some model $\mathfrak{M} \models \mathcal{T}$ such that $\mathfrak{M} \models b(R_i(\tau_i))$ for all $1 \leq i \leq k$.

Lemma 3.1. *Let $\mathcal{T}$ have only infinite models. Suppose, after Skolemization, $\mathcal{T}$ is not disjoint. Then there exists some $\mathcal{T}'$, gotten from $\mathcal{T}$ by addition of a finite number of axioms and constants, that is disjoint and has only infinite models.*

Proof. If $\mathcal{T}$ is not disjoint, then there exists a disjoint set of relations $\mathcal{R} = \{R_i(\tau_i) : 1 \leq i \leq k\}$, which will involve some finite subset $T \subset \mathbb{N} \setminus C$ as arguments, as well as a

forbidden assignment $b \in \{0,1\}^{\mathcal{R}}$, such that every injection $\iota : T \to \mathfrak{M}$ into every model $\mathfrak{M} \models \mathcal{T}$ disagrees with b. We will take $\mathcal{R}$ to be minimal inclusion wise. Then letting $T' \subset T$ be the elements mentioned by $\mathcal{R}' = \mathcal{R} \setminus \{R_k(\tau_k)\}$ there exists a model $\mathfrak{M} \models \mathcal{T}$ and injection $\iota : T' \to \mathfrak{M}$ agreeing with b on $\mathcal{R}'$ (even if T' is empty). Let $E \subset M$ be the (finitely many) elements in the image of ι. Then, for any δ sending the elements in $T \setminus T'$ mentioned by $R_k(\tau_k)$ to $\mathfrak{M} \setminus E$, $R_k(\delta(\tau_k))$ is forced to the opposite of $b(R_k(\tau_k))$. So we add to the principle the constants E and the axioms $R_i(\iota(\tau_i)) = b(R_i(\tau_i))$, for $1 \leq i \leq k-1$. This new principle has at least one infinite model (to wit, $\mathfrak{M}$), and any instantiation $R_k(y_k)$ of R_k mentioning no elements in E is redundant (as its value is forced). We can only repeat this procedure finitely many times, as each relation starts off with some fixed arity, and any relation playing the role of R_k above from then on is only ever instantiated with at least one more constant than it was before. □

In the proof of Lemma 3.1, we showed that if we were in a particularly difficult situation, like a lack of disjointness, then there must be an actionable type of logical redundancy modulo some finite number of axioms, and that this type of redundancy can only be rectified some finitely many times, so eventually we should land in a more favourable situation. The steps below often work similarly - we will find and remove redundancies like $\forall x, y\ P(x,y) \implies P(y,x)$ (the redundancy here being the order implicit in the ordering of the arguments, and later on after transformation, the order implicit in subscript of a variable of a linear program) or that $\forall x, y, z(R(x,y,z) \implies \forall w R(x,y,w))$ (so we did not need the last coordinate). In order to more obviously show finite progress, we first will convert any given SP refutation D of (the n-th CNF translation of) some theory $\mathcal{T}$ into a refutation D' of a related $\mathcal{T}'$. $\mathcal{T}'$ has an expanded vocabulary (but no constants) and will enable us to view the refutation D' as only mentioning relations with distinct parameters, instantiated in order as natural numbers, and crucially, we will have that $|D'| \leq |D|$.

Definition 3.3. Fix some theory $\mathcal{T}$ over some language $\mathcal{L}$. A subset $\mathcal{L}' \subseteq \mathcal{L}$ will be called *sufficient* if, for all literals $R(\tau)$, there is a literal $R'(\tau')$ instantiated from $\mathcal{L}'$ such that $\mathcal{T} \models R(\tau) \Leftrightarrow R'(\tau')$, and the elements mentioned by $R'(\tau')$ are a subset of those mentioned by $R(\tau)$.

By simply replacing literals we get the following fact:

Fact 3.1. Given any SP refutation D refuting $\mathcal{T}_n$ and any sufficient J, there is a D' with $|D'| \leq D$ refuting $\mathcal{T}_n$ and only mentioning variables in J.

Lemma 3.2. *For any theory $\mathcal{T}$ over some language $\mathcal{L}$, and for any set C of forbidden constants, we can find an expanded theory $\mathcal{T}'$ over some expanded language $\mathcal{L}'$, such that relations instantiated without any constants are sufficient.*

Proof. To achieve this we include, for every k-ary relation $R(\mathbf{x}_1, \ldots, \mathbf{x}_k)$, $(|C|+1)^k$ new relations, one for each partial assignment of constants to the parameters. Given some partial assignment $\rho : X_\rho \to C$ defined on some subset $X_\rho \subseteq \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$, we enumerate $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\} \setminus X_\rho$ as $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_{k-|X_\rho|}}$ and include in $\mathcal{T}'$ the axiom

$$R_\rho(\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_{k-|X_\rho|}}) \Leftrightarrow R(\gamma_1, \ldots, \gamma_k)$$

where $\gamma_i = \rho(\mathbf{x}_i)$ if $\mathbf{x}_i \in X_\rho$, otherwise, it is $\mathbf{x}_i$, which is free. □

Note that if we were to name a generic constant and apply the procedure just described we might actually lose disjointness. This happens for example in the PHP, which is disjoint, but if we were to name a constant α and then introduce the relation R_j equivalent to $P_{j\alpha}$, we find that any R_j being set to 1 forces the remaining to 0. However if we were to name one more constant, say β, and say that $P_{\beta\alpha} = 1$, the new relation R effectively disappears outside of $\{\beta\}$. The number of new constants required in naming a constant, forbidding it, and then re-establishing disjointness, is bounded by a universal constant which we call the *cost of disjointness*.

Lemma 3.3. *For any theory $\mathcal{T}$ over some language $\mathcal{L}$, we can find an expanded theory $\mathcal{T}'$ over some expanded language $\mathcal{L}'$, such that only relations with distinct parameters are sufficient.*

Proof. For $k \in \mathbb{N}$ let $\mathcal{E}_k$ be the set of equivalence classes, or partitions, of $[k]$. We define an additional set of relations like

$$\mathcal{L}_1 := \{R_\eta(\mathbf{x}_1, \ldots, \mathbf{x}_{|\eta|}) : R(\mathbf{x}_1, \ldots, \mathbf{x}_k) \in \mathcal{L},\ \eta \in \mathcal{E}_k\}$$

and include in $\mathcal{T}'$ the axioms

$$R_\eta(\mathbf{x}_1, \ldots, \mathbf{x}_{|\eta|}) \Leftrightarrow R(\gamma_1, \ldots, \gamma_k)$$

where $\gamma_i = \mathbf{x}_j$ for the partition j into which i falls. □

Lemma 3.4. *For any theory $\mathcal{T}$ over some language $\mathcal{L}$ equipped with a binary predicate $\prec$ along with the axioms of a total order, we can find an expanded theory $\mathcal{T}'$ over some expanded language $\mathcal{L}'$, such that relations instantiated only in an order consistent with $\prec$ are sufficient.*

For example, as we always consider our countable domains as the natural numbers, the reader could take $\prec$ as just being the standard ordering of $\mathbb{N}$. This lemma is a

generalisation of the discussion following Lemma 2.11, where we show we might as well take subscripts to be unordered.

Proof of Lemma 3.4. Let $\mathcal{S}_k$ be the set of permutations of $[k]$. We define an expanded set of relations like

$$\mathcal{S}(\mathcal{L}) := \{R^{\pi}(\mathbf{x}_1, \ldots, \mathbf{x}_k) : R(\mathbf{x}_1, \ldots, \mathbf{x}_k) \in \mathcal{L},\ \pi \in \mathcal{S}_k\}$$

and include in $\mathcal{T}'$ the axioms

$$R(\mathbf{x}_1, \ldots, \mathbf{x}_k) \Leftrightarrow R^{\pi^{-1}}(\mathbf{x}_{\pi 1}, \ldots, \mathbf{x}_{\pi k})$$

letting us replace any $R(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ with $R^{\pi^{-1}}(\mathbf{x}_{\pi 1}, \ldots, \mathbf{x}_{\pi k})$, where π is the unique permutation bringing the $\mathbf{x}_i$ into agreement with $\prec$, and πk is the application of π to k. □

Note that as we are interested in infinite models it does no harm to include such a $\prec$. Note also that each process in the previous two Lemmas preserves the property gained by the prior - Lemma 3.3 and Lemma 3.4 do not reintroduce any constants, and Lemma 3.4 does not introduce repeated parameters, so we get the following corollary:

Corollary 3.1. *For any theory $\mathcal{T}$ over some language $\mathcal{L}$ equipped with a binary predicate $\prec$ along with the axioms of a total order and any set of named constants C, we can find an expanded theory $\mathcal{T}'$ over some expanded language $\mathcal{L}'$, such that there is a sufficient set of instantiated relations $M_C \subseteq \mathcal{L}'$ where*

1. *nothing in M_C is ever instantiated with any constants in C,*
2. *nothing in M_C is ever instantiated with repeated parameters, and*
3. *all the parameters in M_C are given in order of $\prec$.*

Definition 3.4. Given some relation $R(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ and some subset $X \subseteq \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$, we define the 'unlabeling' to be the formula gotten by universally quantifying over the variables in X:

$$U_X(R(\mathbf{x}_1, \ldots, \mathbf{x}_k)) := \forall (y_i : \mathbf{x}_i \in X) R(y_1, \ldots, y_k)$$

where the y_i for $\mathbf{x}_i \notin X$ are just set to $\mathbf{x}_i$.

Definition 3.5. A theory $\mathcal{T}$ will be called *distilled* if, for every relation R and nonempty X,

$$\mathcal{T}, \mathcal{A} \not\models R(\mathbf{x}_1, \ldots, \mathbf{x}_k) \implies U_X(R(\mathbf{x}_1, \ldots, \mathbf{x}_k)).$$

where $\mathcal{A}$ is any finite number of axioms consistent with $\mathcal{T}$.

Lemma 3.5. *Given some D refuting some theory $\mathcal{T}$, we can produce some D' refuting some distilled theory $\mathcal{T}'$ with $|D'| = |D|$.*

Proof. While $\mathcal{T}$ is not already distilled, we have, for some finite number of axioms $\mathcal{A}$, that

$$\mathcal{T}, \mathcal{A} \models R(\mathbf{x}_1, \ldots, \mathbf{x}_k) \implies \forall \mathbf{x}_l, \ldots, \mathbf{x}_k R(\mathbf{x}_1, \ldots, \mathbf{x}_{l-1}, \mathbf{x}_l, \ldots, \mathbf{x}_k)$$

where we have assumed (purely to simplify the exposition) that the $\mathbf{x}_i$ with $\mathbf{x}_i \in X$ are $\mathbf{x}_l, \ldots, \mathbf{x}_k$ for some l. But this clearly implies

$$\mathcal{T}, \mathcal{A} \models \neg R(\mathbf{x}_1, \ldots, \mathbf{x}_k) \implies \forall \mathbf{x}_l, \ldots, \mathbf{x}_k \neg R(\mathbf{x}_1, \ldots, \mathbf{x}_{l-1}, \mathbf{x}_l, \ldots, \mathbf{x}_k)$$

(as, if for some instantiation of $\mathbf{x}_l, \ldots, \mathbf{x}_k$ there exists a further instantiation of $\mathbf{x}_l, \ldots, \mathbf{x}_k$ rendering $R(x_1, \ldots, x_k)$ true, then any instantiation of $\mathbf{x}_l, \ldots, \mathbf{x}_k$ would have done the same) and that any instance of $R(\mathbf{x}_1, \ldots, \mathbf{x}_{l-1}, \mathbf{x}_l, \ldots, \mathbf{x}_k)$ in $\mathcal{T}$ (and D) can be replaced with some smaller $R'(\mathbf{x}_1, \ldots, \mathbf{x}_{l-1})$, and for any $\mathfrak{M} \models \mathcal{T}$ consistent with $\mathcal{A}$ (of which there are at least one), there is some $\mathfrak{M}' \models \mathcal{T} \cup \mathcal{A}$ of the same cardinality that interprets $R'(\mathbf{x}_1, \ldots, \mathbf{x}_{l-1})$ as true if and only if $\mathfrak{M}$ interpreted any (and all) $R(\mathbf{x}_1, \ldots, \mathbf{x}_{l-1}, \mathbf{x}_l, \ldots, \mathbf{x}_k)$ as true. Of course this process can only be repeated finitely many times, as we permanently reduce the finite arity of one of finitely many relations. □

3.1.2 The lower bound

Definition 3.6. Given some theory $\mathcal{T}$ with constants in $\mathbb{N}$, $n \in \mathbb{N}$, and assignment $\rho = \{R_1(\overline{\tau}_1) \to b_1, \ldots, R_k(\overline{\tau}_k) \to b_k\}$ where the τ_i are instantiated with parameters from $[n]$, we let $h_\rho(\mathcal{T}, n)$ be the injections of $[n]$ into some model $\mathfrak{M}$ of $\mathcal{T}$ that are consistent with ρ.

Definition 3.7. Fix some ambient $n \in \mathbb{N}$. Given some consistent assignment ρ to some set of variables $\{R_1(\overline{\tau}_1) = b_1, \ldots, R_k(\overline{\tau}_k) = b_k\}$, where every τ_i mentions elements in $[n]$, and some subset $\mathcal{N} \subseteq \mathcal{L}$, we define $F_\mathcal{N}(\rho)$ to be the *literals* in $\mathcal{N}$ 'that follow from', or are forced by, ρ:

$$F_\mathcal{N}(\rho) := \{R(\tau) \ : \ R(\tau) \neq R_i(\overline{\tau}_i) \text{ for any } i \text{ but every } \iota \in h_\rho(\mathcal{T}, n) \text{ also satisfies } R(\iota(\overline{x}))\}.$$

where the τ in $R(\tau)$ are instantiated with elements from $[n]$.

We let $F(\rho) := F_{\mathcal{L}}(\rho)$ (where $\mathcal{L}$ is the entire language of $\mathcal{T}$). We will also identify $F_{\mathcal{N}}(\rho)$ with the conjunction of its elements (which is a finite conjunction because we draw the arguments from some finite set $[n]$).

Lemma 3.6. *Let $\mathcal{T}$ be a Skolemized theory over a language $\mathcal{L}$. Let a be an upper bound on the arity of any relation in $\mathcal{L}$ (including the Skolem ones) and d be the number of constants in $\mathcal{T}$. Suppose we have some binary assignment $\rho = \{R_1(\tau_1) \to b_1, \ldots, R_k(\tau_k) \to b_k\}$ of some instantiated relations from the language of $\mathcal{T}$, and where the τ_i all mention elements in $[n]$ but leave at least $d + a + 2$ elements of $[n]$ unmentioned. If $h_\rho(\mathcal{T}, n)$ is nonempty, the assignment*

$$\alpha_\rho(R(\tau)) := \begin{cases} b_i & \text{if } R(\tau) = R_i(\tau_i) \text{ for some } i, \\ 1 & \text{if } R(\tau) \in F(\rho), \\ 0 & \text{if } \neg R(\tau) \in F(\rho), \\ 1/2 & \text{otherwise} \end{cases}$$

is admissible.

Proof. First we show that all the small clauses are satisfied. If every variable $\mathcal{R} = \{R_i(x_i) : 1 \leq i \leq c\}$ in a small clause is set to something binary, then that small clause must be satisfied. This is because there exists at least one model $\mathfrak{M} \models \mathcal{T}$ and some $\iota :\in h_\rho(\mathcal{T}, n)$ such that $R_i(\tau_i)$ is set to the interpretation in $\mathfrak{M}$ of $R_i(\iota(\tau_i))$, and because we were given that ρ was consistent.

If there are two or more elements set to $1/2$ then the clause is already satisfied. So the last real case is if only a single variable v is set to $1/2$. In this case every other variable appearing in the clause is binary, which means its value is forced by every $\iota : [n] \to \mathfrak{M}$ consistent with ρ, for every $\mathfrak{M} \models \mathcal{T}$. If an assignment falsifies such a clause its because v is forced to 0 or 1, and not both because ρ was promised to be consistent - but then $v \in F(\rho)$ should have already been forced to the correct value.

For the Skolem clauses

$$\sum_{i=1}^{n} S(\tau, i) \geq 1$$

there are at least two elements $e_1, e_2 \in [n]$ not appearing in ρ, in τ or as constants in $\mathcal{T}$, as ρ involves at most $d + a + 2$ elements. If τ does not already have a S-witness in

$[n] \setminus \{e_1, e_2\}$ then we must have $S(\tau, e_1)$ and $S(\tau, e_2)$ set to 1/2 - given any $\iota : [n] \to \mathfrak{M}$ consistent with ρ, we simply change $\iota(e_1)$ to point to the S witness of τ in $\mathfrak{M}$. □

Corollary 3.2. *For every $\mathcal{T}$ and $n \in \mathbb{N}$, there is a constant $c_{\mathcal{T}}$ such that any SP refutation D of $\mathcal{T}_n$ must mention at least $n - c_{\mathcal{T}}$ elements.*

Proof. Let $c_{\mathcal{T}} := d + a + 2$, where d and a are as in the setup to the previous Lemma 3.6. Let $E \subset \mathbb{N}$ be all the elements mentioned by D. Let ρ_E be any 'full assignment' - pick a model $\mathfrak{M} \models \mathcal{T}$, ι any injection of the elements E into $\mathfrak{M}$, and for every relation R in the language of $\mathcal{T}$ and every tuple τ of mentioned elements M of the appropriate arity, set $R(\tau)$ to be the interpretation by $\mathfrak{M}$ of $R(\iota(\tau))$. Then these $\rho_E, \mathfrak{M}$, and ι meet the prerequisites of Lemma 3.6 and we receive an admissible point fully integral on all variables contained in E, including every variable appearing in a query in D - but then this admissible point can fall into no slab of D. □

We give an illustrative example for PHP. Note it is already disjoint- the first pigeon roosting or not roosting in the second hole doesn't force the third pigeon to roost or not roost in the fourth hole. Let $\mathcal{R} = \{P_{1,2}, P_{3,4}, \dots P_{2k-1,2k}\}$ be a disjoint set and ρ be anything in $\{0,1\}^{\mathcal{R}}$. Then

$$\alpha_\rho(P_{a,b}) = \begin{cases} \rho(P_{2i-1,2i}) & \text{if } a, b = 2i-1, 2i \text{ for some } 1 \leq i \leq k \\ 0 & \text{if } \rho(P_{a',b}) = 1 \text{ for some } a' \neq a \\ 1/2 & \text{otherwise} \end{cases}$$

is admissible. In this case, $F(\rho)$ are the variables in the middle line, aka the $P_{j,2i}$ for $j \neq 2i-1$ that are forced to 0 by $P_{2i-1,2i}$ being set to 1.

As in the previous Chapter, we would like to find a large hypercube in some set of admissible points. So far we have shown that given some relations instantiated with disjoint parameters we can assume that they can take any truth value independently, however there is still some work to do, as non-hypercube coordinates might depend too much on the assignment to the hypercube ones.

As an example, take the *functional* PHP_4, which asks that every pigeon is assigned to *exactly* one hole (classically the PHP asks only for at least one hole per pigeon). The variables P_{12} and P_{34} can consistently take any of four assignments $\rho \in \{0,1\}^2$. However, the value of $\alpha_\rho(P_{13})$ depends on the value of $\rho(P_{12})$ (it's 0 if P_{12} is set to 1, otherwise it's 1/2). In examples with higher arity, non-hypercube variables may have a nonlinear dependence on multiple hypercube variables. Still in the functional PHP we can find a 'scaled hypercube' - if we set P_{12}, P_{34} to 0 we receive an admissible point (the a_ρ)

where every other variable is set to 1/2, and importantly, which remains admissible for all 4 ways of setting any of P_{12}, P_{34} to 1/2. This is not true for any binary assignment, however - if we had set P_{12} to 1, then P_{13} and P_{14} would be forced to 0, which themselves force P_{12} to 1, preventing us from finding a hypercube. We show now that we can always find an assignment that isn't, in this sense, 'self reinforcing'.

Lemma 3.7. *Let $\mathcal{T}$ be disjoint and distilled. Given any disjoint set of instantiated literals $\mathcal{R} = \{R_1(\overline{x}_1), \dots R_f(\overline{x}_f)\} \subseteq M$, where $M = M_C$ for any C as defined in Corollary 3.1, and the $\overline{x}_i$ althogether mention fewer than $n - c_{\mathcal{T}}$ elements (where $c_{\mathcal{T}}$ is defined in the body of Corollary 3.2), there exists a consistent assignment σ of $\mathcal{R}$ such that for all $2^{|\mathcal{R}|}$ subsets K of $\mathcal{R}$, the assignment $\alpha_{\sigma,K}$ gotten by taking the assignment α_σ from Lemma 3.6 and setting the elements in K to 1/2, that is,*

$$\alpha_{\sigma,K}(v) := \begin{cases} \alpha_\sigma(v) & \text{if } v \notin K \\ 1/2 & \text{otherwise} \end{cases}$$

is admissible.

Proof. Fix any assignment σ of $\mathcal{R}$, necessarily consistent by disjointness. We one by one set the elements of $\mathcal{R}$ in α_σ to 1/2. If we remain admissible no matter how we go about this then the lemma is proven (we can always set any K to 1/2). However if at some point we find we cannot set some $R_i(\overline{x}_i)$ to 1/2 in α_σ whilst remaining admissible, then $R_i(\overline{x}_i)$ is forced by the remaining assignment to its current assignment, that is, $\mathcal{T} \models \sigma(\mathcal{R}_i') \wedge F_M(\sigma) \implies \sigma(R_i(\overline{x}_i))$, where $\mathcal{R}_i'$ contains all the elements in $\mathcal{R}$ aside from $R_i(\overline{x}_i)$ that were not set to 1/2. Letting $\mathcal{R}_i = \mathcal{R} \setminus \{R_i(\overline{x}_i)\}$ and noting (as $\mathcal{R}_i' \subseteq \mathcal{R}_i$) that then $T \models \sigma(\mathcal{R}_i) \wedge F_M(\sigma) \implies \sigma(R_i(\overline{x}_i))$, to prove the Lemma, it would be enough to give an assignment σ such that, for all $1 \leq i \leq f$, $T \not\models \sigma(\mathcal{R}_i) \wedge F_M(\sigma) \implies \sigma(R_i(\overline{x}_i))$.

Suppose we cannot - then for each of the $\{0,1\}^{\mathcal{R}}$ assignments σ to $\mathcal{R}$, there exists at least one $R_i(\overline{x}_i) \in \mathcal{R}$, which we will call *problematic for* σ, such that we have the pair of implications (in $\mathcal{T}$)

$$\sigma \implies F_M(\sigma) \qquad \text{and} \qquad F_M(\sigma) \cup \sigma(\mathcal{R}_i) \implies R_i(\overline{x}_i).$$

(But note that the first implication is anyway always true by definition, and here we have used the sufficiency of M.) Starting from any σ, flip the value of $\sigma(R_i)$ for some R_i problematic for σ, and repeat. Either we eventually find some nonproblematic assignment or we flip some $R \in \mathcal{R}$ twice. In the second case let σ_1 be the working assignment the first time we flip R and σ_{k+1} the assignment the second time. Without loss of generality, $R = R_1(\overline{x}_1)$ was the variable flipped twice and the other variables flipped exactly

once between σ_1 and σ_{k+1} were $R_2(\overline{x}_2)\ldots, R_k(\overline{x}_k)$ in that order. So, to illustrate, $\sigma_1 = \sigma(R_1(\overline{x}_1)), \sigma(R_2(\overline{x}_2))\ldots, \sigma(R_k(\overline{x}_k)), \sigma'$, and $\sigma_3 = \neg\sigma(R_1(\overline{x}_1)), \neg\sigma(R_2(\overline{x}_2))\ldots, \sigma(R_k(\overline{x}_k)), \sigma'$, and $\sigma_{k+1} = \neg\sigma(R_1(\overline{x}_1)), \ldots, \neg\sigma(R_k(\overline{x}_k)), \sigma'$, where $\sigma' := \bigwedge_{i=k+1}^{f} \sigma_1(R_i(\overline{x}_i))$ are the unflipped variables.
Then we have the following set of implications, all apparently consequences of $\mathcal{T}$:

$$\begin{aligned} F_M(\sigma_1) \wedge \sigma_1(\mathcal{R}_1) &\implies R_1(\overline{x}_1) \qquad (3.1)\\ F_M(\sigma_2) \wedge \sigma_2(\mathcal{R}_2) &\implies R_2(\overline{x_2})\\ F_M(\sigma_3) \wedge \sigma_3(\mathcal{R}_3) &\implies R_3(\overline{x_3})\\ &\vdots\\ F_M(\sigma_{k+1}) \wedge \sigma_{k+1}(\mathcal{R}_1) &\implies \neg R_1(\overline{x}_1) \qquad (3.2) \end{aligned}$$

(where we have assumed, for simplicity and without loss of generality, that σ_1 is the all-true assignment). For all $2 \leq i \leq k$ we pick two injections π_i, μ_i of $\overline{x}_i$ into $\mathbb{N}$, and for all $k < i \leq f$, we pick a single injection γ_i of $\overline{x}_i \to \mathbb{N}$. We pick these injections such that the images of all π_i, μ_i, and γ_i are pairwise disjoint. We add said images as constants, add the axioms $R_i(\pi(\overline{x}_i)) \wedge \neg R_i(\mu(\overline{x}_i))$ for all $2 \leq i \leq k$, and the axiom $\sigma'(R_i(\gamma_i(\overline{x}_i)))$ for the remaining $k < i \leq f$. This expanded $\mathcal{T}'$ makes sense by disjointness of $\mathcal{T}$, and in $\mathcal{T}'$ we have

$$R_1(\overline{x}_1) \Leftrightarrow F'(\sigma_1) \qquad \text{and} \qquad \neg R_1(\overline{x}_1) \Leftrightarrow F'(\sigma_{k+1}),$$

where $F'(\sigma_i)$ is the result of taking $F_M(\sigma_i)$ and setting the $\overline{x}_j$ for $j \leq k$ to π_j or μ_j when $\sigma_i(R_j(\overline{x}_j))$ is positive or negative, respectively, and the remaining $\overline{x}_j$, $j > k$, to γ_j. This is because, essentially, $\sigma_1(\mathcal{R}_1)$ in Equation (3.1) becomes a conjunction of axioms, giving the direction $F'(\sigma_1) \implies R_1(\overline{x}_1)$. For the other direction, as the implication $\sigma_1 \implies F_m(\sigma_1)$ was already true in $\mathcal{T}$, after viewing all conjuncts in σ_1 aside from $R_1(\overline{x}_1)$ as axiomatic, we see that $R_1(\overline{x}_1) \implies F'(\sigma_1)$. The right hand side of the 'and', where R_1 is negated, is in like case.

Say $F'(\sigma_1)$ fails to mention R_1 at all. Then, considering $F'(\sigma_1)$ as a formula $F'_{\sigma_1}(\overline{x}_1)$ in the free variables $\overline{x}_1$, we can remove R_1 from $\mathcal{L}$ and replace every instance of $R_1(\overline{y})$ with $F'_{\sigma_1}(\overline{y})$. Now say $F'(\sigma_1)$ does mention R_1 - then it has to mention some $R_1(\overline{y})$ where $\overline{y}$ mentions something in $\mathbb{N}$ outside of $\overline{x}_1$ (as otherwise $\overline{y}$ is a permutation of $\overline{x}_1$, or contains repetitions, both of which are impossible by the assumption of $R_1(\overline{x}_1)$ being a member of the M from Corollary 3.1). If $R_1(\overline{y})$ is mentioned positively, we have

$$R_1(\overline{x}_1) \implies F'(\sigma_1) \wedge U_{A_1}(F_M(\overline{x}_1)) \implies U_{A_1}(F_M(\overline{x}_1))$$

where A_1 is the (nonempty) set of coordinates on which $\overline{y}$ is not contained in $\overline{x}_1$. To summarise, then, $R_1(\overline{y})$ is mentioned negatively in $F'(\sigma_1)$, and as the situation is analogous, $R_1(\overline{y})$ is mentioned positively in $F'(\sigma_{k+1})$. But then

$$\neg R_1(\overline{x}_1) \implies F'(\sigma_{k+1}) \wedge U_{A_{k+1}}(R_1(\overline{x}_1)) \implies U_{A_{k+1}}(R_1(\overline{x}_1)), \text{ and}$$
$$R_1(\overline{x}_1) \implies F'(\sigma_1) \wedge U_{A_1}(\neg R_1(\overline{x}_1)) \implies U_{A_1}(\neg R_1(\overline{x}_1)) \implies U_{A_1}(U_{A_{k+1}}(R_1(\overline{x}_1))))$$
$$\implies U_{A_1 \cup A_{k+1}}(R_1(\overline{x}_1)),$$

where the second to last implication follows from the chain of implications on the line above, contradicting the idea that $\mathcal{T}$ was reduced.

□

Definition 3.8. Given some linear polynomial q in the variables associated with $\mathcal{T}_n$, we say that q is *covered* by a set $B \subseteq [n]$ if every variable appearing with nonzero coefficient in q mentions an element in B, and we say that q has covering number k if every the minimally sized such B has cardinality k.

Here we diverge in notation and nomenclature of the previous chapter - given some function $f(n)$, we say that a query q is $f(n)$-wide if its covering number is at least $f(n)$, and that a query is $f(n)$-narrow if it is not $f(n)$-wide. (Recall that previously we said instead that a polynomial was narrow if it was contained fully in the set, rather than just covered by it.)

Lemma 3.8. *Let $\mathcal{T}$ be some FO principle. Suppose we have some SP refutation D of $\mathcal{T}_n$ where $|D| < n^{1/4}$. Then, if n is large enough, we can find some SP refutation D' of $\mathcal{T}'_m$, where $m \geq n/2$, D' contains only $\nu n^{3/4}$-wide queries for some constant ν, $\mathcal{T}'$ is disjoint, and $|D'| \leq |D|$.*

Proof. Let c be the cost of disjointness as defined in the discussion following Lemma 3.8 and let a be the maximum arity of any relation in the principle under attack. Due to Fact 3.1 we will assume without loss of generality that the variables in D are all in the sufficient M_C as defined in Corollary 3.1, and where initially $C = \emptyset$ (so no constants are forbidden yet).

Let ν be $(2ac)^{-1}$ which as promised is independent of n. Suppose there exists a query $q \in D$ that is $\nu n^{3/4}$ narrow. Then there is some set $B \subseteq [n]$ with $|B| \leq \nu n^{3/4}$ touching every variable in q. We forbid this set B, updating the subscript C in M_C as $B \cup C$, and re-establish disjointness as in Lemma 3.1. As every variable in q was touched by B we drop the maximum arity of a variable appearing in q by at least one, and so we do this at most $an^{1/4}$ times (as it was claimed that there at most $n^{1/4}$ many $q \in D$). So

we consume in this way at most $an^{1/4} \cdot c\nu n^{3/4} \leq ac\nu \cdot n = n/2$ elements, the remaining $q \in D$ (if there are any) refuting $\mathcal{T}'_{(1-ac\nu)n}$ and all $\nu n^{3/4}$-wide. □

We finally arrive at our Theorem.

Theorem 3.2. *Let $\mathcal{T}$ be some sanitised theory with only infinite models. Then any SP refutation D of $\mathcal{T}_n$ has size $\Omega(n^{1/4})$.*

Proof. Assume the opposite. Then, for any fixed $\kappa > 0$ of our choosing, for large enough n, there is D with $|D| < \kappa(n/4)^{1/4}$ refuting $\mathcal{T}_n$. By Lemma 3.8, we can find some D' refuting $\mathcal{T}'_m$, with $\mathcal{T}'$ disjoint, $|D'| \leq |D|$, $m \geq n/2$, every query $\nu n^{3/4}$-wide, and ν a positive constant independent of n.
Let a be the maximum arity of any relation in the language of $\mathcal{T}'$. We initialise an empty hypercube H and process the queries $q_1, \ldots, q_{|D|}$ in any order. When processing q_i we pick $\nu a^{-1}\sqrt{n/4}$ disjoint variables, disjoint from everything in H - we can always do this as there are at most $\kappa(n/4)^{1/4} \cdot \nu a^{-1}(n/4)^{1/2} = \kappa\nu a^{-1}(n/4)^{3/4}$ variables in H, mentioning then at most $\kappa\nu(n/4)^{3/4}$ elements, and q is $\nu(n/3)^{3/4}$-wide (and we will be careful to choose $\kappa \leq 1$). Then, by Lemma 3.7, we find a hypercube of size 2^H covered by D' where every $q \in D'$ has effective width at least $\nu a^{-1}\sqrt{n/4}$. Then, by Lemma 2.2, every $q \in D'$ rules out at most

$$\frac{2^{|H|}}{\sqrt{\nu a^{-1} \cdot \sqrt{n/4}}} = (\nu/a)^{-1/2}\frac{2^{|H|}}{(n/4)^{1/4}}$$

elements in 2^H, and so by choosing $\kappa < \sqrt{\nu/a}$ we see that altogether they kill strictly less than

$$|D'| \cdot (\nu/a)^{-1/2}\frac{2^{|H|}}{(n/4)^{1/4}} < \kappa(\nu/a)^{-1/2}2^{|H|} < 2^{|H|}$$

points, contradicting Lemma 1.2. □

3.1.3 Conclusions

We began this section by showing that we can assume the ability to find certain useful structure in the solution space of $\mathcal{T}_n$, for *any* principle $\mathcal{T}$ admitting only infinite models. We then took advantage of this structure to show a polynomial sized lower bound for Stabbing Planes. However, this raises two questions:

Firstly, what is the situation of the matching upper bound? In [28], a similar lower bound to Theorem 3.2 is given - there the authors show that, if $\mathcal{T}$ admits only infinite models (but at least one), then every SA refutation of $\mathcal{T}_n$ has polynomial rank. They

then continue to completely classify every relevant principle $\mathcal{T}$, by then showing that if $\mathcal{T}$ has no models at all, finite or infinite, then it has constant rank SA refutations. This complementary upper bound is currently missing for SP. Corollary 2.1 suggests the mirrored upper bound cannot be the same for SP, however, the principle concerned (the SPHP) is not obviously the result of the transformation of some FO principle.

Secondly, what is the situation for other geometric proof systems, such as CP? As SA simulates CP in terms of size, Theorem 3.2 actually also applies to CP - but there a polynomial sized lower bound is much less meaningful. However, the results of Section 3.1.1 are quite general, being model-theoretic in nature. We harnessed them against SP, but one could imagine their application to a proof system like CP, perhaps replacing the final application of the antichain method with the method of the protection lemma used in [15] and elsewhere, giving a similar lower bound, but for CP rank (instead of size).

3.2 Sum of Squares

Let $\mathcal{T}$ be a consistent FO formula and $\mathcal{T}_n$ be the propositional formula claiming $\mathcal{T}$ has a model of finite size n, always presumed to be contradictory. A natural pseudodistribution appearing often in the literature is a 'pseudomodel-counting' distribution, where, for each clause C, we count the number of size-n substructures of any model of $\mathcal{T}$, labeled by elements in $[n]$, consistent with C, and then normalise to generate a probability. The 'pseudo' prefix is important, as there no actual models of any finite size n by assumption. (The term pseudomodel here is used vaguely but is defined precisely in Definition 3.13.)

For example, the pseudodistribution used in [71] to give a a $\Omega(n^{1/2-\epsilon})$ lower bound for the LOP (any $\epsilon > 0$) is an example of a pseudomodel-counting argument - every clause can be identified with a partial order, and the pseudodistribution used counts the number of linear extensions of this partial order - that is, it counts pseudomodels. For illustration, the degree-2 clause $P_{1,2}P_{3,4}$ would receive the valuation $1/4$ - exactly half of all linear orderings of $[n]$ place 1 before 2, and out of these, exactly half place 3 before 4.

A further example is in [29] which gives uses a pseudomodel counting argument for a different principle - the PHP. This time, a clause mapping some pigeons to holes invectively is valued proportionately to the number of matchings on the domain that are consistent with this mapping (this valuation is in fact exactly the valuation defined and used later in Lemma 4.1, albeit in a different context). And perhaps the example closest to the theme of this section is [28], where it is shown that this generic counting distribution gives, for all relevant FO principles $\mathcal{T}$, a polynomial in n SA rank lower bound for $\mathcal{T}_n$.

In this section, we begin the task of bringing this sort of lower bound to SOS. We first show in the discussion following Definition 3.13 that this generic pseudodistribution immediately works at any degree for all parts of the principle, apart from matrices generated from the existential Skolem clauses.

Then we appeal to well-known results regarding symmetry and SOS. For every principle $\mathcal{T}$, $\mathcal{T}_n$ as a set of clauses or linear inequalities is actually symmetric with respect to perumutation of the labels in $[n]$ (apart from perhaps a universally bounded number of constants which we can ignore - see Lemma 3.2 from the previous chapter, for example), In [70, 73] it is shown that to a large extent we can expect this symmetry of $\mathcal{T}_n$ to actually be reflected in the structure of any SOS refutation of $\mathcal{T}_n$ (see Corollary 3.3 for the precise result).

The main thrust of this section is the recognition that as a consequence of this symmetry, pseudomodel counting amounts to a counting of embedding from small pseudomodels into larger ones. Essentially, as we can identify clauses in the same orbit (with respect to some permutation group $\mathcal{G}$), we can identify the models consistent with those clauses (aka, the objects we count to produce the pseudodistribution) according to the same symmetry. Then, elements in $[n]$ subject to the symmetry of $\mathcal{G}$ are free to 'float around', and we find ourselves counting partially labelled embeddings (this is the content of Lemma 3.10). We use this observation to reduce the PSDness of the final remaining matrix into a condition on a summation of these counts (Corollary 3.4).

Let $\mathcal{P}(X, \leq d)$ denote the subsets of X of size at most d, $\mathcal{P}(X, d)$ those of size exactly d, and $\mathcal{P}(X)$ the usual power set of X. The element in the mth row and ith column of a matrix M is denoted $M(m, i)$. In this section our matrices will be always indexed by sets of variables (aka. square-free monomials) and will obey $M(\alpha, \beta) = M(\alpha', \beta')$ whenever $\alpha \cup \beta = \alpha' \cup \beta'$ so sometimes we will index them as if they are one dimensional. Given some fixed ambient 'relational formal variables' $A, B, \ldots$ with arities $m_A, m_B, \ldots$, we let V_n be the set of variables $\{A_{a_1,\ldots,a_{m_A}}, \ldots, B_{b_1,\ldots,b_{m_B}}, \ldots\}$ where the subscripts range over $[n]$. $\mathbb{R}[V_n]$ is the set of formal polynomials with variables in V_n. Given a product of variables $I = \prod_{x \in J \subseteq V_n} x$ and a polynomial $p \in \mathbb{R}[V_n]$ we let $p[I]$ denote the coefficient in p before I. We will often in this way identify a polynomial with its vector of coefficients.

3.2.1 Symmetry

We need some standard group-theoretic definitions and concepts. From now on, we will let an element $\sigma \in S_{[n]}$ act on monomials in $\mathbb{R}[\mathcal{V}_n]$ like

$$\sigma(P_{l_1 r_1} \cdots P_{l_m r_m}) \to P_{\sigma(l_1)\sigma(r_1)} \cdots P_{\sigma(l_m)\sigma(r_m)}. \tag{3.3}$$

Here a permutation σ is viewed as an injective function.
(For expository purposes we have assumed only one binary relation in our language - this action very straightforwardly applies to the more general case.)

Definition 3.9. Some vector x, indexed by monomials in atomic variables, is said to be *symmetric with respect to* some permutation group $\mathcal{G}$ if, for all $\sigma \in \mathcal{G}$ and α indexing x:

$$x(\sigma(\alpha)) = x(\alpha).$$

Definition 3.10. Fix some permutation group $\mathcal{G}$. The *orbit* of β, denoted $\mathcal{O}(\beta)$, is

$$\mathcal{O}(\beta) := \{\pi(\beta) \ : \ \pi \in \mathcal{G}\}.$$

If some x is symmetric with respect to some $\mathcal{G}$, every coordinate in the same orbit is given the same value, and we may index vectors by orbits.

3.2.2 Representation theory of the symmetric group

Definition 3.11. A *representation* (V, θ) of some group G over some vector space V is a group homomorphism $\theta : G \to GL(V)$ (the group of invertible linear operators from V to itself).

Note that Equation (3.3) gives such a representation of permutation groups over $\mathbb{R}[\mathcal{V}_n]$, where some permutation is sent to the corresponding permutation matrix. As we are only interested in this action we henceforth omit θ and assume $V \subseteq \mathbb{R}[\mathcal{V}_n]$.

Definition 3.12. A representation V is said to be *irreducible* if it has no proper non-trivial subrepresentations, i.e., the only $W \subseteq V$ fixed by the action of G is the zero space and V itself.

The representation theory of the symmetric group is well understood, and it reveals that this representation of S_k (or any representation of any finite group over a field of zero characteristic) breaks up into a direct sum of irreducible components.

For S_k these components are indexed by so called Young tableaux. These tableau have a *shape* which is a partition of k - a nonincreasing list of numbers $\lambda = (\lambda_1, \ldots, \lambda_m)$ that sum up to k (denoted $\lambda \vdash k$). A tableau of shape λ is a filling of λ by some elements of $[k]$. An example tableau for the partition $(3, 2, 2) \vdash 7$ is

1	3	5
2	4	
6	7	

.

A tableau is called *standard* if it is increasing along rows and down columns (like this example tableau).

So we have something like

$$\mathbb{R}[\mathcal{V}_n] = \bigoplus_{\lambda \vdash k} V_\lambda$$

with each V_λ being a further sum of irreducible components indexed by tableau of shape λ: $V_\lambda = \sum_{i=1}^{m_\lambda} V_\lambda^i$ for some multiplicities m_λ, and $\oplus$ is the direct sum.

For some space U acted on by a group G, let U^G be the subspace of U fixed by the action of G:

$$U^G := \{u \in U : \ g \cdot u = u \ \text{ for all } g \in G\}.$$

The following is summarising almost verbatim from [73]. We fix any group G. Let $(S_\lambda)_{\lambda \in \Lambda}$ be an enumeration of inequivalent irreducible representations of G. Let $\mathrm{Hom}_G(U, W)$ be the vector space of 'intertwining operators': $\mathbb{R}$-linear maps $\phi : U \to W$ such that $\phi(g \cdot u) = g \cdot \phi(u)$ for all $u \in U$.

Define

$$V_\lambda = \mathrm{span}\{\phi(s) \ : \ \phi \in \mathrm{Hom}_G(S_\lambda, V), \ \ s \in S_\lambda\} \tag{3.4}$$

$$W_{s_\lambda} := \{\phi(s_\lambda) \ : \ \phi \in \mathrm{Hom}_G(S_\lambda, V)\} \subseteq V_\lambda. \tag{3.5}$$

Theorem 3.3 (Theorem A.8 from [73]). *Let V be a finite-dimensional G-invariant subspace of $\mathbb{R}[\mathsf{x}]/\mathcal{I}$ with isotypic decomposition $V = \bigoplus_{\lambda \in \Lambda} V_\lambda$ and corresponding multiplicities m_λ. For each $\lambda \in \Lambda$, fix a non-zero element $s_\lambda \in S_\lambda$. Let $\mathsf{b}_1, \ldots, \mathsf{b}_{m_\lambda}$ be a basis for the subspace W_{s_λ}. Define for each $\lambda \in \Lambda$*

$$Y_{ij}^\lambda = \mathit{sym}(\mathsf{b}_i \cdot \mathsf{b}_j) := \frac{1}{|G|} \sum_{\sigma \in G} \sigma(\mathsf{b}_i) \cdot \sigma(\mathsf{b}_j)$$

for $i, j \in [m_\lambda]$. Suppose $\mathsf{p} \in \mathbb{R}[\mathsf{x}]/\mathcal{I}$ is invariant under the action of G and is V-SOS (a sum of squared polynomials all living in V). Then there exist $m_\lambda \times m_\lambda$ PSD matrices Q^λ such that

$$\mathsf{p} = \sum_{\lambda \in \Lambda} \sum_{i,j \in [m_\lambda]} Q_{ij}^\lambda Y_{ij}^\lambda.$$

Theorem 3.4 (Corollary 2.6 in [73]). *Let $\Lambda := \{\lambda \vdash n \ : \ \lambda \geq_{\mathrm{lex}} (n - 2d, 1^{2d})\}$. Suppose $\mathsf{p} \in \mathbb{R}[V]$ is $S_{[n]}$-invariant and d-SOS (a sum of squared polynomials, all of degree up to d). For each partition $\lambda \vdash n$, fix a tableau τ_λ of shape λ and choose a vector space basis $\{\mathsf{b}_1^{\tau_\lambda}, \ldots, \mathsf{b}_{m_\lambda}^{\tau_\lambda}\}$ for W_{τ_λ}. Then for each partition $\lambda \in \Lambda$, there exists a $m_\lambda \times m_\lambda$ PSD matrix Q_λ such that*

$$\mathsf{p} = \sum_{\lambda \in \Lambda} \mathrm{tr}(Q_\lambda \, Y^{\tau_\lambda}).$$

Theorem 3.4 is useful for the following reason, given in [70]. Suppose we have some linear operator $E : \mathbb{R}[V] \to \mathbb{R}$ which is symmetric with respect to $S_{[n]}$, and we are interested in the sign of $E[g^2]$ for some degree-d polynomial g. Due to the symmetry of E we have $E[g^2] = E[\frac{1}{n!} \sum_{\sigma \in S_{[n]}} (\sigma(g^2))]$. The polynomial on the right hand side is $S_{[n]}$-invariant

and a d-SOS, and therefore we can invoke Theorem 3.4 and find

$$E\left[\sum_{\sigma\in S_{[n]}}(\sigma(g^2))\right] = E\left[\sum_{\lambda\in\Lambda}\mathrm{tr}(Q_\lambda Y^{\tau_\lambda})\right] = \sum_{\lambda\in\Lambda} E\left[\mathrm{tr}(Q_\lambda Y^{\tau_\lambda})\right].$$

Now, again due to the symmetry of E, the Y matrices can be 'desymmetrized' and written as $B_{\tau_\lambda}^\top B_{\tau_\lambda}$ (for some matrices B_{τ_λ}), and due to the PSDness of the Q matrices we can also write $Q_\lambda = K_\lambda^\top K_\lambda$. We then end up with

$$E\left[\mathrm{tr}(Q_\lambda Y^{\tau_\lambda})\right] = \sum E\left[\left(\sum c_i \mathsf{b}_i^{\tau_\lambda}\right)^2\right]$$

i.e., we are applying the operator to a sum of squared polynomials in the subspaces W_{τ_λ}, which are symmetric with respect to (at least) permutations of the head of τ_λ, which is of size at least $n-2d$. In summary,

Lemma 3.9. *In order to check whether or not $E[p^2]\geq 0$ for some degree d polynomial p and some linear operator $E:\mathbb{R}[V]\to\mathbb{R}$ symmetric with respect to $S_{[n]}$, it suffices to only check polynomials symmetric with respect to permutations of all but $2d$ elements of $[n]$.*

Corollary 3.3. *Let M be some matrix symmetric with respect to $S_{[n]}$. In order to show that $x^\top Mx\geq 0$ for every x mentioning at most d elements, it is enough to show that $x^\top Mx\geq 0$ for every x symmetric with respect to the permutation group fixing $[2d]$.*

3.2.3 The pseudodistribution

Fix some first order sentence $\mathcal{T}$ having at least one infinite model and no finite models. (If there are finite models then we either do not produce contradictions, and if there are no infinite models there are constant degree Sherali-Adams [23] refutations, and therefore the same for SOS, as SOS simulates SA [10].)

Let $k,d\in\mathbb{N}$, where $k\geq d$. Let A_k denote the set of monomials mentioning up to k elements. (These are in the relational variables associated with $\mathcal{T}$ in the sense defined in Section 1.1.) A lower bound on the number of elements mentioned gives a degree lower bound, as the monomials of degree up to d can mention up to md elements, m being the maximum arity of a relation in our language, which we assume to be a constant independent of n.

Definition 3.13. A *$\mathcal{T}$-substructure* of size $n\in\mathbb{N}$ is any size-n substructure of some (by assumption infinite) model satisfying $\mathcal{T}$, labeled with elements from $[n]$. A size-n *$\mathcal{T}$-substructure* is called a A *$\mathcal{T}$-pseudomodel* if it is maximal with respect to the partial

order $\preceq$ defined on size-n $\mathcal{T}$-substructures like

$$A \preceq B \text{ if for all relations } R \text{ in our language and all tuples } \overline{a} = (a_1, \ldots, a_m) \in [n]^m,$$
$$R^A(\overline{a}) \implies R^B(\overline{a}),$$

where $R^A(\overline{a})$ is the interpretation by A of $R^A(\overline{a})$.
We say that a pseudomodel *satisfies* a monomial if it assigns true to all variables appearing in that monomial. Given some monomial C mentioning only elements in $[n]$, let $\Lambda_X(C)$ be the set of pseudomodels labeled by X and satisfying C, and $\Lambda_n(C) = \Lambda_{[n]}(C)$. $\Lambda_X = \Lambda_X(\varnothing)$ denotes all X-labeled pseudomodels. Finally, for some pseudomodel σ, we let $E(\sigma)$ denote the atoms true in σ.

We define our generic pseudodistribution to be

$$L[C] := \frac{|\Lambda_n(C)|}{|\Lambda_n|}. \tag{3.6}$$

(So the pseudodistribution is just the probability that a pseudomodel chosen at random from $|\Lambda_n|$ is consistent with C.)

We perhaps have gotten ahead of ourselves by calling L a pseudodistribution already. In order to do so legitimately we must show that it satisfies the conditions of Definition 1.10. For the PSDness of the moment matrix $M(\alpha, \beta) = L[\alpha \cup \beta]$, denote the vectors $\sigma_\rho \in \{0,1\}^{A_d}$ for each $\rho \in \Lambda_n$ by $\sigma_\rho(\alpha) = 1$ if and only if $\rho \in \Lambda_n(\alpha)$ and 0 otherwise. Then

$$M = \sum_{\rho \in \Lambda_n} \sigma_\rho \sigma_\rho^\top$$

is obviously PSD.

In the unary case we have to show the PSDness of the the localising matrix arising from some instantiation $\overline{x}, \overline{y}$ of the universal clause Equation (1.2):

$$C = \sum_{i \in V} (1 - S_i(\overline{x}_1, \ldots, \overline{x}_i, \overline{y}_i)) + F(\overline{x}, \overline{y}) \geq 1$$

With the characteristic vectors σ_μ defined as before, this localising matrix is by definition

$$Y = \sum_\mu \left(\sum_{i \in V} (1 - S_i^\mu(\overline{x}_1, \ldots, \overline{x}_i, \overline{y}_i)) + F^\mu(\overline{x}, \overline{y}) \right) \sigma_\mu \sigma_\mu^\top$$

where, given some formula or relation $F(\overline{v})$, $F^{\mu}(\overline{v})$ is 0/1 depending on whether $F(\overline{v})$ is false in μ or true, respectively. Each multiplier in front of each $\sigma_{\mu}\sigma_{\mu}^{\top}$ is nonnegative (due to μ being a substructure), and so the matrix Y is PSD, for any degree. For the same reason, the localising matrices of the small clauses are always positive semidefinite.

Flags. The notation here is borrowed from [74], with minimal modifications to suit it for our purposes.

Fix some ambient FO formula $\mathcal{T}$ and some $k \in \mathbb{N}$. A *k-flag* $\rho = (D_{\rho}, M_{\rho}, \theta_{\rho})$ is a $\mathcal{T}$-substructure M_{ρ} with domain D_{ρ}, unlabeled apart from k distinguished elements pointed at by the injective $\theta_{\rho} : [k] \to M_{\rho}$. (So a flag is basically a partially labeled $\mathcal{T}$-substructure.) $\mathcal{F}_n^k$ denotes all k-flags of size n, and $\mathcal{F}_k$ denotes the k-flags of any size. We let $|\rho| = |D_{\rho}|$. Given X a subset of the ground set of M_{ρ} not in the image of θ, $\rho|_X$ is the sub-$\mathcal{T}$-substructure induced by X. A flag embedding α of ρ_l into ρ_r (both k-flags) is a model embedding $M_{\rho_l} \to M_{\rho_r}$ respecting the labelings from $[k]$: $\alpha(\theta_{\rho_l}(i)) = \theta_{\rho_r}(i)$ for all $i \in [k]$. We relate two flags by $\lambda \subseteq \mu$ if there exists a flag embedding from λ into μ, and $\lambda \subset \mu$ if $\lambda \subseteq \mu$ and $\lambda \neq \mu$. Given some k-flag ρ of size $n > k$, we define the *instantiations* $\mathcal{I}_{\rho}$ of ρ to be all the $\mathcal{T}$-substructures gotten by labeling the unlabeled elements of ρ with the remaining labels $[n] \setminus [k]$. Finally, for $\rho \in \mathcal{F}_n^k$, we define $\Lambda^{\rho}(C) \subseteq \mathcal{I}_{\rho}$ to be the instantiations of ρ that satisfy C.

Define $\mathcal{H}(\alpha, \rho)$ to be the set of flag embeddings of α into ρ and let $H(\alpha, \rho) := |\mathcal{H}(\alpha, \rho)|$.

Definition 3.14. Let $\mathcal{M}_{\rho}$ be the (PSD) matrix defined by $\mathcal{M}_{\rho}(\alpha, \beta) = |\Lambda^{\rho}(\alpha \cup \beta)|$.

Lemma 3.10. *Let $C := \exists i\ S(\overline{x}, i)$ be any Skolem clause and fix any tuple $\overline{g} \supseteq \overline{x}$ of some distinguished elements of $[n]$. The PSDness of a localizing matrix of the Skolem clause C (*Skolem matrix *for short) follows from the PSDness of a linear combination of the form $\sum_{\rho \in \mathcal{F}_n^k} a_{\rho}\mathcal{M}_{\rho}$, where each $a_{\rho} \in \mathbb{Z}$ is the number of witnesses in ρ of the existential claim generating the Skolem clause, minus one.*

Proof. By relabelling we will assume $\overline{g}$ is of the form $(1, 2, \ldots, |\overline{g}|)$.

We are concerned with the sum

$$|\Lambda_n|^{-1} \sum_{\alpha,\beta \in A_d} x(\alpha)x(\beta) \left(\sum_{i=1}^{n} L\left[\alpha \cup \beta \cup S(\overline{x}, i)\right] - L\left[\alpha \cup \beta\right] \right). \tag{3.7}$$

Now, due to Corollary 3.3, we can assume that $x(\alpha) = x(\beta)$ for any two α, β differing only by a permutation of $[n] \setminus [g]$. Then this can be expressed as

$$\sum_{\alpha,\beta \in \mathcal{F}_{2g}^g} x(\alpha)x(\beta) \sum_{\substack{\sigma \in \mathcal{I}_{\alpha} \\ \sigma' \in \mathcal{I}_{\beta}}} \left(\sum_{i=1}^{n} L\left[\sigma \cup \sigma' \cup S(\overline{x}, i)\right] - L\left[\sigma \cup \sigma'\right] \right) \tag{3.8}$$

where we have let $g := 2|\overline{g}|$. We now show that the two terms in the subtraction together are of the form claimed.

Lemma 3.11. *Let $\alpha, \beta \in \mathcal{F}^g_{2g}$. We have*

$$\sum_{\substack{\sigma \in \mathcal{I}_\alpha \\ \sigma' \in \mathcal{I}_\beta}} L\left[\sigma \cup \sigma'\right] = (n')! \sum_{\rho \in \Lambda^g_n} H(\alpha, \rho) H(\beta, \rho) = \sum_{\rho \in \Lambda^n_g} \mathcal{M}_\rho.$$

Proof. The notation in this proof appears quite dense, however the structure of the proof is much more simple than the notation does suggest, and we describe it informally first. Here α and β are partially labelled pseudomodels of size $2g$, where the exactly the first g elements are labelled. The leftmost term

$$\sum_{\substack{\sigma \in \mathcal{I}_\alpha \\ \sigma' \in \mathcal{I}_\beta}} L\left[\sigma \cup \sigma'\right]$$

is counting the number of fully labelled pseudomodels of size n consistent with any 'filling in' of the unlabelled elements, aka. an instantiation of, α and β. Let ι_α and ι_β be any such pair of instantiations, and let $\rho \in \Lambda_n(\iota_\alpha \cup \iota_\beta)$ be a (fully labelled) pseudomodel consistent with both ι_α and ι_β. This consistency of ρ with ι_α and ι_β can already be seen as a 'static' embedding of ι_α and ι_β into ρ, in the sense that, if ι_α mentions some set $E \subseteq [n]$ (so $|E| = 2g$), ρ, when projected onto the elements in E, will contain ι_α as a substructure.

For a concrete example the reader may take the LOP. Suppose $g = 2$ and let $\alpha = P_{1,2}P_{1,3}P_{1,4}$. Now ρ could be the standard linear ordering

$$1 < 2 < 3 < 4 < 5 < 6 < \ldots < n,$$

of $[n]$ as natural numbers, which is certainly consistent with ι_α, and if one imagines this linear ordering as a transitively closed directed path, the claw α can be found embedded as the (noninduced) subgraph in the projection of ρ onto $\{1, 2, 3\}$, which is the lowest part of the order. (This is where the terms of the form $I|_{\mathrm{img}(e)}$ come from in the main body of the proof to come.)

But note that, for any distinct pair $\{i, j\} \subseteq [n] \setminus [2]$, $P_{1,2}P_{1,i}P_{1,j}$ is also consistent with ρ, and is also isomorphic to ι_α with respect to the symmetric group $S_{[n]\setminus[g]}$. So all $(n-2)(n-3)$ embeddings of $P_{1,2}P_{1,i}P_{1,j}$ into ρ are isomorphic to α, and they act to 'carry around' the static embedding of ι_α into the prefix $\{1, 2, 3\}$ over all embeddings of the unlabelled originator $\alpha \in \mathcal{F}^g_{2g}$ into the labelled version of ρ.

Then the term $(n-g)!$ comes from the unlabelling of ρ, by identifying pseudomodels in Λ_n^g that differ only by some $\pi \in S_{[n]\setminus[g]}$ - for example, the static embedding of the α into the bottom of ρ is really the same embedding of $\alpha' = P_{1,2}P_{1,n}P_{1,(n-1)}$ into

$$\rho' = 1 < 2 < n < n-1 < n-2 < \ldots < 4 < 3,$$

and so is overcounted by each of the $(n-g)!$ permutations of $[n] \setminus [g]$.

Now, more formally, every pair $e \in H(\alpha, \rho), e' \in H(\beta, \rho)$ of embeddings into ρ corresponds exactly with $(n-g)!$ pairs $\sigma \in \mathcal{I}_\alpha, \sigma' \in \mathcal{I}_\beta$ counted in $\sum_{\substack{\sigma \in \mathcal{I}_\alpha \\ \sigma' \in \mathcal{I}_\beta}} L\left[\sigma \cup \sigma'\right]$, which are gotten by labelling everything outside of $[g]$. Namely, for each of the $(n-g)!$ instantiations $I \in \mathcal{I}_\rho$, we have $I \in \Lambda^\rho((I|_{\mathrm{img}(e)}) \cup (I|_{\mathrm{img}(e')}))$. So $\sum_{\substack{\sigma \in \mathcal{I}_\alpha \\ \sigma' \in \mathcal{I}_\beta}} |\Lambda^\rho(\sigma \cup \sigma')| = (n-g)!H(\alpha, \rho)H(\beta, \rho)$ and

$$\sum_{\alpha,\beta \in \mathcal{F}_{2g}^g} x(\alpha)x(\beta) \sum_{\substack{\sigma \in \mathcal{I}_\alpha \\ \sigma' \in \mathcal{I}_\beta}} |\Lambda^\rho(\sigma \cup \sigma')| = (n-g)! \sum_{\alpha,\beta \in \mathcal{F}_{2g}^g} x(\alpha)x(\beta)H(\alpha, \rho)H(\beta, \rho).$$

□

Lemma 3.12.

$$\sum_{\substack{\sigma \in \mathcal{I}_\alpha \\ \sigma' \in \mathcal{I}_\beta}} \sum_{i=1}^{n} L\left[\sigma \cup \sigma' \cup S(\overline{x}, i)\right] = (n-g)! \sum_{\rho \in \Lambda_n^g} \omega_\rho H(\alpha, \rho)H(\beta, \rho) = \sum_{\rho \in \Lambda_g^n} \omega_\rho \mathcal{M}_\rho$$

where ω_ρ is the number of potential witnesses in ρ.

Proof. Similar to the previous lemma. The new term ω_ρ comes from the fact that now, each pair of embeddings $e \in H(\alpha, \rho), e' \in H(\beta, \rho)$ is not just counted once for each of the $(n-g)!$ instantiations of ρ, but additionally once more for each assignment of the index i appearing in $S(\overline{x}, i)$ to each of the ω_ρ potential existential witnesses in ρ. □

□

The coordinates indexing vectors need not correspond to pseudomodels - for example, in the case of the TLNP, our vectors are currently indexed by partial orders, although we count over total orders. The next lemma will address this.

Lemma 3.13. *Fix $\mathcal{L} = \sum_{\rho \in \Lambda_n} a_\rho \mathcal{M}_\rho$ for some coefficients a_ρ. Given vectors $x, v \in \mathbb{R}^{A_k}$, nonzero only on elements in A_d (for some d) and symmetric with respect to some permutation group $\mathcal{G}$, we can produce vectors x', v' where*

1. x', v' *are symmetric with respect to* $\mathcal{G}$,
2. x', v' *are only nonzero on pseudomodels of size at most* d, *and*
3. $x^\top \mathcal{L} v = (x')^\top \mathcal{L} v'$.

Proof. Given any size-d subset $Y \subseteq [n]$ and any $\sigma \in \Lambda_Y$, define the clause

$$U_\sigma := \prod_{e \in E(\sigma)} e.$$

If every nonzero coordinate in x and y is of the form U_σ for some σ we are done - otherwise, fix some α and assume (without loss of generality, as $\mathcal{L}$ is symmetric and we can take the transpose of $x^\top \mathcal{L} y$) that $x(\mathcal{O}(\alpha)) \neq 0$.

Let X_α be the set of elements mentioned by α and $\Gamma(\alpha) := \Lambda_{X_\alpha}(\alpha)$ be the size-$|\alpha|$ pseudomodels labeled with X_α that respect α.

We show that a generic clause α is really shorthand for a collection of submodels. More formally, for any β,

$$\Lambda^\rho(\alpha \cup \beta) = \bigcup_{\sigma \in \Gamma(\alpha)} \Lambda^\rho(U_\sigma \cup \beta).$$

The inclusion from right to left comes from noting that α can be embedded into any pseudomodel respecting U_σ for any $\sigma \in \Gamma(\alpha)$, and the inclusion from left to right comes from noting that every $\sigma \in \Lambda^\rho(\alpha)$ is in $\Lambda^\rho(U_{\sigma'})$ with σ' being σ restricted to X_α. The sets on the right are pairwise disjoint (as any distinct $\sigma, \sigma' \in \Lambda_{X_\alpha}$ disagrees on the status of some edge in X_α, due to maximality according to $\prec$) so we even get

$$|\Lambda^\rho(\alpha \cup \beta)| \; = \sum_{\sigma \in \Gamma(\alpha)} |\Lambda^\rho(U_\sigma \cup \beta)|. \tag{3.9}$$

This gives, for any fixed β,

$$\sum_{\sigma \in \Gamma(\alpha)} \mathcal{L}(U_\sigma, \beta) = \sum_{\sigma \in \Gamma(\alpha)} \sum_{\rho \in \Lambda_n} a_\rho |\Lambda^\rho(U_\sigma \cup \beta)| = \sum_{\rho \in \Lambda_n} a_\rho |\Lambda^\rho(\alpha \cup \beta)| = \mathcal{L}(\alpha, \beta). \tag{3.10}$$

We are going to add $x(\alpha)$ to all the $x(U_\sigma), \sigma \in \Gamma(\mathcal{O}(\alpha))$ and then set $x(\mathcal{O}(\alpha))$ to zero. The two vectors reflecting this are:

$$y(\beta) = \begin{cases} 0 & \text{if } \beta \in \mathcal{O}(\alpha) \\ x(\beta) & \text{otherwise.} \end{cases} \qquad z(\gamma) = \begin{cases} \chi_\sigma \cdot x(\alpha) & \text{if } \gamma = U_\sigma \text{ for some } \sigma \\ 0 & \text{otherwise.} \end{cases} \tag{3.11}$$

where $\chi_\sigma := |\{\alpha' \in \mathcal{O}(\alpha) \ : \ \sigma \in \Gamma(\alpha')\}|$. We will use

$$\chi_{\pi\cup\sigma} = |\left\{\alpha' \in \mathcal{O}(\alpha) \ : \ \pi \cup \sigma \in \Gamma(\alpha')\right\}| = |\left\{\alpha' \in \mathcal{O}(\alpha) \ : \ \sigma \in \Gamma(\pi^{-1}(\alpha'))\right\}| = \chi_\sigma. \tag{3.12}$$

We claim that

$$(y+z)^\top \mathcal{L} v = x^\top \mathcal{L} v.$$

The term $y^\top \mathcal{L} v$ is exactly $x^\top \mathcal{L} v$ missing any pairs where $x(\mathcal{O}(\alpha))$ appears:

$$y^\top \mathcal{L} v = x^\top \mathcal{L} v - x(\alpha) \left(\sum_{\alpha' \in \mathcal{O}(\alpha)} \sum_{\beta} v(\beta) \mathcal{L}(\alpha', \beta) \right).$$

The second term $z^\top \mathcal{L} v$ can be calculated from the definitions:

$$\begin{aligned} z^\top \mathcal{L} v &= \sum_{\beta,\gamma} z(\gamma) v(\beta) \mathcal{L}(\beta,\gamma) = x(\alpha) \sum_{\sigma} \sum_{\beta} \chi_\sigma \mathcal{L}(U_\sigma, \beta) \\ &= x(\alpha) \left(\sum_{\substack{\alpha' \in \mathcal{O}(\alpha) \\ \sigma \in \Gamma(\alpha')}} \sum_{\beta} x(\beta) \mathcal{L}(U_\sigma, \beta) \right) \\ &= x(\alpha) \left(\sum_{\alpha' \in \mathcal{O}(\alpha)} \sum_{\beta} x(\beta) \mathcal{L}(\alpha', \beta) \right). \end{aligned}$$

The second equality follows from the fact that $z(\gamma)$ is nonzero if and only if $\gamma = U_\sigma$. The third followed from Equation (3.10) and the discussion preceding it. We can repeat this until we get vectors of the appropriate form.

It remains to show that the symmetry under $\mathcal{G}$ is preserved. v has not been changed, so we only need to check $(y+z)$. And it is easy to see that each of the vectors y, z are still symmetric under $\mathcal{G}$: for y, a single orbit is set to 0 (and it is equal to x otherwise). For z we need to show that $z(\gamma) = z(\pi(\gamma))$ for any $\pi \in \mathcal{G}$:

$$z(\pi(\gamma)) = \begin{cases} \chi_\sigma \cdot x(\alpha) & \text{if } \pi(\gamma) = U_\sigma \text{ for some } \sigma \\ 0 & \text{otherwise.} \end{cases} = \begin{cases} \chi_\sigma \cdot x(\alpha) & \text{if } \gamma = \pi^{-1}(U_\sigma) \text{ for some } \sigma \\ 0 & \text{otherwise.} \end{cases}$$

and we finish by noting that $\pi^{-1}(U_\sigma) = U_{\pi^{-1}\cup\sigma}$ and that $\chi_\sigma = \chi_{\pi^{-1}\cup\sigma}$. □

Let Λ_n^i, $0 \leq i \leq n$, denote the pseudomodels in Λ_n where there are i potential witnesses. Then, appealing to Lemma 3.12, Equation (3.8) becomes

$$\begin{aligned}
&\sum_{i=0}^{n}(i-1)\sum_{\rho\in\Lambda_n^i}\left(\sum_{\alpha}x_\alpha H(\alpha,\rho)\right)^2 \\
&=\sum_{i=0}^{n}(i-1)\sum_{\rho\in\Lambda_n^i}\left(\sum_{\alpha,\beta}x_\alpha x_\beta H(\alpha\cup\beta,\rho)\right) \\
&=\sum_{\alpha,\beta}x_\alpha x_\beta\sum_{i=0}^{n}(i-1)\sum_{\rho\in\Lambda_n^i}H(\alpha\cup\beta,\rho).
\end{aligned}$$

So letting $T_i := \sum_{\rho\in\Lambda_n^i} M_\rho$, we have the following:

Corollary 3.4. *Let $g = g(n)$ be some function of n. If $\sum_{i=0}^{n}(i-1)T_i$ is positive semidefinite whenever the T_i are indexed by elements in Λ_{2g}^g, then the principle considered has SOS degree $\Omega(g)$.*

3.2.4 Conclusions

In this section, we investigated a generic pseudodistribution appearing more specifically throughout the literature, and we reduced the complex problem of showing the PSDness of the several localising matrices to the PSDness of just a single type of matrix - a comprehensible weighted summation of outer products, where the vectors forming these outer products simply count homomorphisms from smaller structures into larger structures, and the weights are just the number of witnesses in that larger structure (minus one). Furthermore, these vectors are well studied [64, 74], and they enjoy many algebraic properties that one imagines could be used to complete this result by showing the PSDness of this single type of matrix.

Unfortunately we did not achieve this completion in this chapter, and this task is left as an exercise for future researchers. We note that any results in this vein could not be exactly the same dichotomy as found in [28], where existence of infinite models alone and enough shows hardness of the principle. This is because, as we note below in Section 4.5, the PHP already has constant degree in a proof system strictly weaker than SOS, despite having infinite models, yet the LOP retains much more hardness [71]. An accomplishment of this result would have to notice a further difference between these two principles.

We are also left wondering, again, about the corresponding upper bound. As SOS simulates SA, we can appeal to the upper bound in [28] and know any principle admitting

no models at all has a constant degree SOS refutation. But this does not work for the PHP for exactly the reasons just stated - it does have an (infinite) model. Exactly what model theoretic properties create hardness or easiness for SOS is at the time of writing is still completely unknown.

Chapter 4

Sherali-Adams and Binary Encodings

In the previous chapter, we investigated the proof complexity of propositional contradictions generated uniformly from FO sentences by the mechanism given by Riis in 2001 [77]. But we could consider different translations from the FO to the propositional. For example, in this chapter, we study what happens if we opt to use a more 'concise' encoding of existential demands. So for the Pigeonhole Principle, instead of insisting that the ith pigeon goes into a hole by saying

$$\sum_{j=1}^{n} P_{i,j} \geq 1$$

we have instead $\lceil \log n \rceil$ many variables $P'_{i,b}$, $1 \leq b \leq \lceil \log n \rceil$, any binary setting of which will point to the hole that the ith pigeon supposedly goes in to. We call the first encoding considered so far the *unary* encoding, and the second one the *binary* encoding.

For the unary encoding of the Pigeonhole Principle and the Least Ordering Principle, it is known that linear rank is required for refutations in SA, although both admit refutations of polynomial size [29, 75]. We prove that the binary encoding of the Pigeonhole Principle requires exponentially-sized SA refutations, whereas the binary encoding of the Least Ordering Principle admits logarithmic rank, polynomially-sized SA refutations. We continue by considering a refutation system between SA and Sum Of Squares, which we call 'SA +Squares'. In this system, the unary encoding of the Least Ordering Principle requires linear rank while the unary encoding of the Pigeonhole Principle becomes constant rank.

4.1 Introduction

Following Riis ([77] and summarised in Section 1.1.1), for any tuple $\overline{x}$, it is typical to encode the existential demand of their being at least one least one witnesses satisfying some FO formula in longhand, with a big disjunction of the form $S_{\overline{x},1} \vee \ldots \vee S_{\overline{x},n}$, asking explicitly that the first element of the domain serves as a witness, or that the second element serves as a witness, etcetera. So for example, in the pigeonhole clause $\exists i\ P_{ij}$, j is the sole member of the tuple $\overline{x}$, and after fixing some finite domain size n this becomes $\bigvee_{i\in[n]} P_{ij}$ (that the witness becomes before or after the tuple in the subscript is only superficial). This we designate the *unary encoding*. If the existential demand in question to be transformed along the lines of Section 1.1.1, the arity of the tuple $\overline{x}$ is the number of universally quantified variables preceding the existentially quantified variable on which it might depend.

As recently investigated in [12, 13, 26, 34, 48, 60], we might opt instead to encode the existence of such witnesses *succinctly* by the use of a *binary encoding*. Briefly, instead of the n many $S_{\overline{x},1} \vee \ldots \vee S_{\overline{x},n}$, we have $\log n$ many variables $S_{\overline{x},1}, \ldots, S_{\overline{x},\log n}$, any binary setting of which implicitly pointing to a proposed witness in $[n]$; whereas in the unary encoding a solitary true literal tells us which is the witness. In this chapter we generally assume, without loss of much generality, that n is a power of 2. Cases where n is not a power of 2 are handled in the binary encoding by explicitly forbidding possibilities (which involves only linearly many clauses and bits).

More specifically, in the context of Section 1.1.1, the binary encoding includes, for every Skolem relation $S_{\overline{x}}$ appearing in the principle, $\lceil \log n \rceil$ many binary variables $\{S_{\overline{x},\alpha} : \alpha \in [\lceil \log n \rceil]\}$. Now the binarization exhibits the following differences from the unary (and recall that at this point the tuples in these clauses have been instantiated with elements in some $[n]$)

- We omit the Skolem clauses (as they exist implicitly due to the fact that any setting of the S_α defines some potential witness w),
- any (unquantified) clause $\left(\bigvee_{i\in V_1} E_i(\overline{r}_i)\right) \vee \left(\bigvee_{i\in V_2} \neg F_i(\overline{s}_i)\right)$ (E_i, F_i relational variables and r_i, s_i tuples of the appropriate arity) remains unchanged, and
- any (unquantified) clause $\left(\bigwedge_{i\in V} S_i(\overline{x}, y_i)\right) \to \left(\bigvee_{i\in V_1} E_i(\overline{r}_i)\right) \vee \left(\bigvee_{i\in V_2} \neg F_i(\overline{s}_i)\right)$ (E_i, F_i relational variables and r_i, s_i tuples of the appropriate arity and y_i free) becomes

$$\left(\bigvee_{i\in V} \bigvee_{j=1}^{\lceil \log n \rceil} S_{i,\overline{x},j}^{1-y_{i,j}}\right) \vee \left(\bigvee_{i\in V_1} (E_i)_{\overline{r}_i}\right) \vee \left(\bigvee_{i\in V_2} \neg (F_i)_{s_i}\right)$$

where, for any variable x, x^1 denotes x and x^0 denotes $\neg x$, and $y_{i,j}$ is the jth bit in the binary representation of y_i. The notation is perhaps cumbersome, but it may be read as 'either some setting of a bit of the S_i is different to the corresponding y_i, or all the S_i are indeed pointing to the y_i and the remaining clause should hold'.

Combinatorial principles encoded in binary are interesting to study for Resolution-type systems since they still preserve whatever 'inherent hardness' of the combinatorial principle while giving a more succinct propositional representation. In certain cases this leads to significant lower bounds more directly than with the unary case [13, 26, 34, 60].

The binary encoding, however, implicitly enforces an at-most-one constraint at the same time as it does at-least-one. When some wide existential disjunction $S_{\overline{x},1} \vee \ldots \vee S_{\overline{x},n}$ of the unary encoding is translated to constraints for an ILP it becomes $S_{\overline{x},1} + \ldots + S_{\overline{x},n} \geq 1$. Were we to insist that $S_{\overline{x},1} + \ldots + S_{\overline{x},n} = 1$ then we encode immediately also the at-most-one constraint. We paraphrase this variant as being (the unary) *encoding with equalities* or 'SA-with-equalities'.

In [29] it was proved that the SA rank of the unary encoding the Pigeonhole Principle and Least Ordering Principles is $n-2$ (where n is the number of pigeons and elements in the poset, respectively). It is known that SA polynomially simulates Resolution (see e.g. [29]) and it follows there is a polynomially-sized refutation in SA of the Least Ordering Principle. That there is a polynomially-sized refutation in SA of the Pigeonhole Principle is noted in [75].

In this chapter we consider the binary encodings of the Pigeonhole Principle and the Least Ordering Principle as ILPs. We additionally consider their (unary) encoding with equalities. We first prove that the binary encoding of the Pigeonhole Principle requires exponential size in SA. We then prove that the (unary) encoding of the Least Ordering Principle with equalities has SA rank 2 and polynomial size. This allows us to prove that the binary encoding of the Least Ordering Principle has SA rank at most $2 \log n$ and polynomial size. The divergent behaviour of these two combinatorial principles is surprising – while the Least Ordering Principle becomes easier for SA in the binary encoding (in terms of rank), the Pigeonhole Principle becomes harder (in terms of size). Such variable behaviour has been observed for the Pigeonhole Principle in Resolution, where the binary encoding makes it easier for treelike Resolution (in terms of size) [26].

We continue by considering a refutation system we call SA+Squares which (in terms of power) lays somewhere in between SA and Sum of Squares. SA+Squares appears as Static LS_+ in [44] In this system one can always assume the non-negativity of (the linearisation of) any squared polynomial. In contrast to our system SA-with-equalities, we see that the rank of the unary encoding of the Pigeonhole Principle is 2, while the

rank of the Least Ordering Principle is linear. We prove this by showing a certain moment matrix is positive semidefinite. Our rank results for the unary encoding can be contrasted in Table 4.1.

unary case	SA	SA-with-equalities	SA+Squares
PHP	linear	linear	constant
OP	linear	constant	linear

binary case	SA
PHP	exponential
OP	polynomial

unary case	SA	SA-with-equalities	SA+Squares
PHP	[29]	[29]	[44]
OP	[29]	Theorem 4.3	Theorem 4.5

binary case	SA
PHP	Theorem 4.1
OP	Corollary 4.3

TABLE 4.1: Rank based complexity for the unary encoding in different systems (on the top) and size based complexity for the binary encoding (on the bottom). The lower table shows where the corresponding result is proved.

4.2 Preliminaries

A *term* is a conjunction of propositional literals. Let us now consider principles which are expressible as first-order formulae, with no finite models, in Π_2-form, i.e. as $\forall\vec{x}\exists\vec{w}\varphi(\vec{x},\vec{w})$ where $\varphi(\vec{x},\vec{w})$ is a formula built on a family of relations $\vec{R}$. For example the *Least Ordering Principle*, which states that a finite partial order has a minimal element, is one of such principles. Its negation can be expressed in Π_2-form as:

$$\forall x,y,z\exists w\ \neg R(x,x)\wedge(R(x,y)\wedge R(y,z)\rightarrow R(x,z))\wedge R(x,w).$$

This can be translated into a unsatisfiable CNF using a unary encoding of the witness, as shown below alongside the binary encoding.

$$\begin{array}{c} \mathsf{OP}_n : \underline{\textit{Unary encoding}} \\ \overline{P}_{i,i} \qquad \forall i \in [n] \\ \overline{P}_{i,j} \vee \overline{P}_{j,k} \vee P_{i,k} \qquad \forall i,j,k \in [n] \\ \overline{S}_{i,j} \vee P_{i,j} \qquad \forall i,j \in [n] \\ \bigvee_{i \in [n]} S_{i,j} \qquad \forall j \in [n] \end{array} \qquad \begin{array}{c} \mathsf{OP}_n : \underline{\textit{Binary encoding}} \\ \overline{P}_{i,i} \qquad \forall x \in [n] \\ \overline{P}_{i,j} \vee \overline{P}_{j,k} \vee P_{i,k} \qquad \forall i,j,k \in [n] \\ \bigvee_{i \in [\log n]} S_{i,j}^{1-a_i} \vee P_{j,a} \qquad \forall j,a \in [n] \\ \text{where } a_1 \dots a_{\log n} = \text{bin}(a) \end{array}$$

Note that we placed the witness in the Skolem variables $S_{i,x}$ as the first argument and not the second, as we had in the introduction. This is to be consistent with the the standard formulation of OP in the proof complexity literature. A more traditional form of the (unary encoding of the) OP_n has clauses $\bigvee_{i \in [n]} P_{i,j}$ which are consequent on $\bigvee_{i \in [n]} S_{i,j}$ and $\overline{S}_{i,j} \vee P_{i,j}$ (for all $i \in [n]$), as was told in Section 1.1.1.

Indeed, one can see how to generate a binary encoding of C from any combinatorial principle C expressible as a first order formula in Π_2-form with no finite models. Exact details can be found in Definition 4 in [26].

As a second example we consider the *Pigeonhole Principle* which states that a total mapping from $[m]$ to $[n]$ has necessarily a collision when m and n are integers with $m > n$. The negation of its relational form for $m = n + 1$ can be expressed as a Π_2-formula like

$$\forall x, y, z \exists w \; \neg P(x, 1) \wedge (P(x, z) \wedge P(y, z) \to x = y) \wedge P(x, w)$$

where the constant 1 represents the object that is among the $[n+1]$ but not among the $[n]$. Its usual unary and binary propositional encoding are:

$$\begin{array}{c} \mathsf{PHP}_n^m : \underline{\textit{Unary encoding}} \\ \bigvee_{j=1}^{n} P_{i,j} \qquad \forall i \in [m] \\ \overline{P}_{i,j} \vee \overline{P}_{i',j} \qquad \forall i \neq i' \in [m], j \in [n] \end{array} \qquad \begin{array}{c} \mathsf{PHP}_n^m : \underline{\textit{Binary encoding}} \\ \bigvee_{j=1}^{\log n} P_{i,j}^{1-a_j} \vee \bigvee_{j=1}^{\log n} P_{i',j}^{1-a_j} \\ \forall a \in [n], i \neq i' \in [m] \\ \text{where } a_1 \dots a_{\log n} = \text{bin}(a) \end{array}$$

where 1 no longer appears now m and n are explicit.

When we consider the Sherali-Adams r-lifts of, e.g., the Least Ordering Principle, we will identify terms of the form $Z_{P_{i,j} \wedge \overline{S}_{i',j'} \wedge \dots}$ as $P_{i,j}\overline{S}_{i',j'} \dots$. Thus, we take the subscript

and use overline for negation and concatenation for conjunction. This prefigures the multilinear notation we will revert to in Section 4.5, but the reader should view for now $P_{i,j}\overline{S}_{i',j'}\ldots$ as a single variable (in a linear program, or a proof system based on linear programming) and not a multilinear monomial. Finally, we wish to discuss the encoding of the Least Ordering Principle and Pigeonhole Principle as ILPs *with equality*. For this, we take the now familiar unary encoding, but instead of translating the wide clauses from $\bigvee_{i\in[n]} S_{i,x}$ to $S_{1,x}+\ldots+S_{n,x}\geq 1$, we instead use $S_{1,x}+\ldots+S_{n,x}=1$. This makes the constraint at-least-one into exactly-one (which is a priori enforced in the binary encoding). A reader who would like a specific reading of the following Lemma should consider the Least Ordering Principle as the combinatorial principle under discourse.

4.3 The lower bound for the binary Pigeonhole Principle

In this section we study the inequalities derived from the binary encoding of the Pigeonhole principle, whose axioms we remind the reader of now. BinPHP^m_n has, for each two distinct pigeons $i\neq i'\in[m]$ and each hole $a\in[n]$, the axiom $\sum_{j=1}^{\log n}\omega_{i,j}^{(1-a_j)}+\sum_{j=1}^{\log n}\omega_{i',j}^{(1-a_j)}\geq 1$, where $a_1\ldots a_{\log n}$ is the binary representation of a. We first prove a certain SA rank lower bound for a version of the binary PHP, in which only a subset of the holes is available.

Lemma 4.1. *Let $H\subseteq[n]$ be a subset of the holes and let us consider $\mathsf{BinPHP}^m_{|H|}$ where each pigeon can go to a hole in H only. Any SA refutation of $\mathsf{BinPHP}^m_{|H|}$ involves a term that mentions at least $|H|$ pigeons.*

Proof. We get a valuation v from a partial matching in an obvious way. That is, if a pigeon i is assigned to hole a, whose representation in binary is $a_1\ldots a_{\log n}$, then we set each $\omega_{i,j}^{a_j}$ to a_j. We say that a product term $P=\prod_{j\in J}\omega_{i_j,k_j}^{b_j}$ mentions the set of pigeons $M=\{i_j : j\in J\}$. Let us denote the number of available holes by $n':=|H|$. Every product term that mentions at most n' pigeons is assigned a value $v(P)$ as follows. The set of pigeons mentioned in M is first extended arbitrarily to a set M' of exactly n' pigeons. $v(P)$ is then the probability that a matching between M' and H taken uniformly at random is consistent with the product term P. In other words, $v(P)$ is the number of perfect matchings between M' and H that are consistent with P, divided by the total, $(n')!$. Obviously, this value does not depend on how M is extended to M'. Also, it is symmetric, i.e. if π is a permutation of the pigeons, $v\left(\prod\omega_{i_j,k_j}^{b_j}\right)=v\left(\prod\omega_{\pi(i_j),k_j}^{b_j}\right)$.

All lifts of axioms of equality $\omega_{j,k}+\neg\omega_{j,k}=1$ are automatically satisfied since a matching consistent with P is consistent either with $P\omega_{j,k}^{b}$ or with $P\omega_{j,k}^{1-b}$ but not with both, and

thus

$$v(P) = v\left(P\omega_{j,k}^{b}\right) + v\left(P\omega_{j,k}^{1-b}\right).$$

Regarding the lifts of the disequality of two pigeons $i \neq j$ in one hole, that is, the inequalities coming from the only clauses in $\mathsf{BinPHP}_{|H|}^{m}$, it is enough to observe that it is consistent with any perfect matching, i.e. at least one variable on the LHS is one under such a matching. Thus, for a product term P, any perfect matching consistent with P will also be consistent with $P\omega_{i,k}^{1-b_k}$ or with $P\omega_{j,k}^{1-b_k}$ for some k. □

4.3.1 The ordinary Pigeonhole Principle

The proof of the size lower bound for the $\mathsf{BinPHP}_{n}^{n+1}$, which we are about to give, is then by a standard random restriction argument, combined with the rank lower bound above. Assume, without loss of generality, that n is a power of two. For the random restrictions $\mathcal{R}$, we consider the pigeons one by one and with probability $1/4$ we assign the pigeon uniformly at random to one of the holes still available. We first need to show that the restriction is "good" with high probability, i.e. neither too big nor too small. The former is needed so that in the restricted version we have a good lower bound, while the latter will be needed to show that a good restriction coincides well with any reasonably big term, in the sense that they have in common a sufficiency of pigeons.

We will make use of the following version of the Chernoff Bound as it appears in [67].

Lemma 4.2 (Theorems 4.4 and 4.5 in [67])**.** *Let $X_1, X_2, \ldots, X_n$ be independent 0/1 random variables with $Pr[X_i = 1] = p_i$. Let $X = \sum_{i=1}^{n} X_i$ and $\mu = E[X]$. Then, for every δ, $0 < \delta \leq 1$, the following bound holds*

$$Pr[X \geq (1+\delta)\mu] \leq e^{\frac{-\mu\delta^2}{3}}$$

and similarly

$$Pr[X \leq (1-\delta)\mu] \leq e^{\frac{-\mu\delta^2}{3}}.$$

Lemma 4.3. *If $|\mathcal{R}|$ is the number of pigeons (or holes) assigned by $\mathcal{R}$, the probability that $|\mathcal{R}| > \frac{3(n+1)}{8}$ is at most $e^{-\frac{(n+1)}{48}}$.*

Proof. We use the Chernoff Bound from Lemma 4.2. We have $p_i = \frac{1}{4}$ (and thus $\mu = \frac{n+1}{4}$) and $\delta = \frac{1}{2}$. Thus, the probability the restriction assigns more than $\frac{3(n+1)}{8}$ pigeons to holes is at most $e^{-(n+1)/48}$. □

We first prove that any given wide product term, i.e. a term that mentions a constant fraction of the pigeons, survives the random restrictions with exponentially small probability.

Lemma 4.4. *Let P be a product term that mentions at least $\frac{n+1}{2}$ pigeons. The probability that P does not evaluate to zero under the random restrictions is at most $\left(\frac{5}{6}\right)^{n/16}$ (for n large enough).*

Proof. We will desire $|\mathcal{R}| \leq \frac{3(n+1)}{8}$ to ensure that at least $\frac{5(n+1)}{8}$ holes remain unused in $\mathcal{R}$ (for n large enough). This will involve the probability $e^{-(n+1)/48}$ from Lemma 4.3.

A further application of the Chernoff Bound from Lemma 4.2 ($\mu = \frac{n+1}{8}$, $\delta = -\frac{1}{2}$) gives the probability that fewer than $\frac{n+1}{16}$ pigeons mentioned by P are assigned by $\mathcal{R}$ is at most $e^{-(n+1)/96}$.

For each of these assigned pigeons the probability that a single bit-variable in P belonging to the pigeon is set by $\mathcal{R}$ to zero is at least $\frac{1}{5}$. This is because when $\mathcal{R}$ sets the pigeon, and thus the bit-variable, there were at least $\frac{5(n+1)}{8}$ holes available, while at most $\frac{n+1}{2}$ choices set the bit-variable to one. The difference – which will be a lower bound on the number of holes available setting the selected bit to 0 – is $\frac{n+1}{8}$ which when divided by $\frac{5(n+1)}{8}$ (to normalise the probability) gives $\frac{1}{5}$. Thus P survives under $\mathcal{R}$ with probability at most $e^{-(n+1)/48} + e^{-(n+1)/96} + \left(\frac{4}{5}\right)^{(n+1)/16} < \left(\frac{5}{6}\right)^{n/16}$. □

Theorem 4.1. *Any SA refutation of the* BinPHP_n^{n+1} *has to contain at least* $\left(\frac{7}{6}\right)^{n/16}$ *terms.*

Proof. Assume, towards a contradiction, that there is a smaller refutation. We wish to argue that there is a random restriction with $|\mathcal{R}| \leq \frac{3(n+1)}{8}$ that evaluates to zero all terms that mention at least $\frac{n+1}{2}$ pigeons. There are at most $\left(\frac{7}{6}\right)^{n/16}$ such terms so an application of the union-bound together with Lemma 4.3 and Lemma 4.4 gives a probability that some term mentioning at least $\frac{n+1}{2}$ pigeons does not evaluate to zero of

$$\left(\frac{5}{6}\right)^{n/16} \times \left(\frac{7}{6}\right)^{n/16} + e^{-(n+1)/48} < 1.$$

Now we apply the random restriction which we know must exist to leave no terms mentioning at least $\frac{n+1}{2}$ pigeons in an SA refutation of the binary $\mathrm{PHP}_{n'}^{m'}$, where $m' > n' \geq \frac{5(n+1)}{8}$. However, since $n' > \frac{n+1}{2}$, this contradicts Lemma 4.1. □

Corollary 4.1. *Any SA refutation of the* BinPHP_n^{n+1} *must have size* $2^{\Omega(n)}$.

4.3.2 The weak Pigeonhole Principle

We now consider the so-called weak binary PHP, BinPHP_n^m, where m is potentially much larger than n. The weak unary PHP_n^m is interesting because it admits (significantly) subexponential-in-n refutations in Resolution when m is sufficiently large [16]. It follows that this size upper bound is mirrored in SA. We will see here that the weak binary BinPHP_n^m remains almost-exponential-in-n for minimally sized refutations in SA. In this weak binary case, the random restrictions $\mathcal{R}$ above do not work, so we apply quite different restrictions $\mathcal{R}'$ that are as follows: for each pigeon select independently a single bit uniformly at random and set it to 0 or 1 with probability of $1/2$ each.

Lemma 4.5. *A product term P that mentions n' pigeons does not evaluate to zero under $\mathcal{R}'$ with probability at most $e^{-n'/2\log n}$.*

Proof. For each pigeon mentioned in P, the probability that the bit-variable present in P is set by the random restriction is $\frac{1}{\log n}$, and if so, the probability that the bit-variable evaluates to zero is $\frac{1}{2}$. Since this happens independently for all n' mentioned pigeons, the probability that they all survive is at most $\left(1-\frac{1}{2\log n}\right)^{n'}$. □

Lemma 4.6. *The probability that $\mathcal{R}'$ fails to have, for each $k \in [\log n]$ and $b \in \{0,1\}$, at least $\frac{m}{4\log n}$ pigeons with the kth bit set to b, is at most $e^{-n/48\log n}$.*

Proof. We apply the Chernoff Bound of Lemma 4.2 to deduce that for each bit position k, $1 \le k \le (\log n)$ and a value b, 0 or 1, the probability that there are fewer than $\frac{m}{4\log n}$ pigeons for which the kth bit is set to b is at most $e^{-m/24\log n}$. This uses $\mu = \frac{m}{2\log n}$ and $\delta = -\frac{1}{2}$. Since $m > n$, by the union bound, the probability that this holds for some position k and some value b is at most $(2\log n)e^{-m/24\log n} \le e^{-n/48\log n}$. □

In order to conclude our result, we will profit from a graph-theoretic treatment of Hall's Marriage Theorem [46]. Suppose G is a finite bipartite graph with bipartitions X and Y, then an X-saturating matching is a matching which covers every vertex in X. For a subset W of X, let $N_G(W)$ denote the neighbourhood of W in G, i.e. the set of all vertices in Y adjacent to some element of W.

Theorem 4.2 ([46] (see Theorem 5.1 in [86]))**.** *Let G be a finite bipartite graph with bipartitions X and Y. There is an X-saturating matching if and only if for every subset W of X, $|W| \le |N_G(W)|$.*

Corollary 4.2. *Any SA refutation of the BinPHP_n^m, $m > n$, has to contain at least $e^{n/32\log^2 n}$ terms.*

Proof. Assume for a contradiction that there is a refutation with fewer than $e^{n/32\log^2 n}$ product terms. We first show, by an application of the probabilistic method standard in proof complexity, that there exists a random restriction evaluating all terms that mention at least $\frac{n}{4\log n}$ pigeons to zero while satisfying the condition of Lemma 4.6. Using a union bound and Lemma 4.5 we upper bound the probability this fails to happen as $e^{-n/8\log^2 n} \cdot e^{n/32\log^2 n} + e^{-n/48\log n} < 1$, so such a random restriction $\mathcal{R}'$ does exist.

Then, $\mathcal{R}'$ leaves at least $\frac{m}{4\log n}$ pigeons of each type (k, b), i.e. the kth bit of the pigeon is set to b. Recalling $m \geq n$, we now pick a set of pigeons S that has $(*)$ precisely $\frac{n}{4\log n}$ pigeons of each type and thus is of size $n/2$.

We will give an evaluation of the restricted principle which contradicts the claim that the original object was a refutation. We evaluate any product term P that mentions at most $\frac{n}{4\log n}$ pigeons by first relabeling the mentioned pigeons, injectively, using the labels of pigeons in S while preserving types, which we can do due to property $(*)$, and then giving it a value as before. That is, by taking the probability that a perfect matching between S and some set of $n/2$ holes consistent with the random restriction, is consistent with P.

To finish the proof, we need to show that such a set of $n/2$ holes exists, that is, such a matching exists. But this follows trivially from Theorem 4.2 as every pigeon has $n/2$ holes available, so at least the same applies to any set of pigeons. □

4.4 The SA rank upper bound for Ordering Principle with equality

Let us remind ourselves of the Ordering Principle in both unary and binary.

$$\underline{\mathsf{OP} : \textit{Unary encoding}}$$

$$\neg v_{i,i} \qquad \forall i \in [n]$$

$$\neg v_{i,j} \vee \neg v_{j,k} \vee v_{i,k} \qquad \forall i, j, k \in [n]$$

$$\neg w_{i,j} \vee v_{i,j} \qquad \forall i, j \in [n]$$

$$\bigvee_{i \in [n]} w_{i,j} \qquad \forall j \in [n]$$

$$\underline{\mathsf{BinOP_n} : \textit{Binary encoding}}$$

$$\neg \nu_{i,i} \qquad \forall i \in [n]$$

$$\neg \nu_{i,j} \vee \neg \nu_{j,k} \vee \nu_{i,k} \qquad \forall i, j, k \in [n]$$

$$\bigvee_{i \in [\log n]} \omega_{i,j}^{1-a_i} \vee \nu_{j,a} \qquad \forall j, a \in [n]$$

$$\text{where } a_1 \dots a_{\log n} = \text{bin}(a)$$

Note that we placed the witness in the variables $w_{i,x}$ as the first argument and not the second, as we had in the introduction. This is to be consistent with the $v_{i,j}$ and

the standard formulation of OP as the least, and not greatest, number principle. A more traditional form of the (unary encoding of the) OP has clauses $\bigvee_{i\in[n]} v_{i,j}$ which are consequent on $\bigvee_{i\in[n]} w_{i,j}$ and $\neg w_{i,j} \vee v_{i,j}$ (for all $i \in [n]$).

In SA we wish to discuss the encoding of the Ordering Principle (and Pigeonhole Principle) as ILPs *with equality*. For this, we take the unary encoding but instead of translating the wide clauses (e.g. from the OP) from $\bigvee_{i\in[n]} w_{i,x}$ to $w_{1,x} + \ldots + w_{n,x} \geq 1$, we instead use $w_{1,x} + \ldots + w_{n,x} = 1$. This makes the constraint at-least-one into exactly-one (which is a priori enforced in the binary encoding). A reader favouring a specific example may consider the Ordering Principle as the combinatorial principle of the following lemma.

Lemma 4.7. *Let* C *be any combinatorial principle expressible as a first order formula in* Π_2*-form with no finite models. Suppose the unary encoding of* C *with equalities has an SA refutation of rank* r *and size* s*. Then the binary encoding of* C *has an SA refutation of rank at most* $r \log n$ *and size at most* s*.*

Proof. We take the SA refutation of the unary encoding of C with equalities of rank r, in the form of a set of inequalities, and build an SA refutation of the binary encoding of C of rank $r \log n$, by substituting terms $w_{x,a}$ in the former with $\omega_{x,1}^{a_1} \cdots \omega_{x,\log n}^{a_{\log n}}$, where $a_1 \ldots a_{\log n} = \mathrm{bin}(a)$, in the latter. $\neg w_{x,a}$ is substituted by $1 - \omega_{x,1}^{a_1} \cdots \omega_{x,\log n}^{a_{\log n}}$. Variables $v_{x,a}$ and $\neg v_{x,a}$ are substituted by $\nu_{x,a}$ and $1 - \nu_{x,a}$, respectively.

It remains to show we can build the translation of the SA with equalities axioms in the binary case from the true axioms of the binary case. Axioms from the binary case that involve only variables ν_{x_a} appear perfectly reproduced. Axioms of the form

$$\sum_{a\in[n]:a_1\ldots a_{\log n}=\mathrm{bin}(a)} \omega_{x,1}^{a_1} \cdots \omega_{x,\log n}^{a_{\log n}} = 1$$

follow from the equalities of negation Equation (1.11). Finally, axioms of the form $\omega_{x,1}^{a_1} \ldots \omega_{x,\log n}^{a_{\log n}} \leq \nu_{x,a}$, can also be built since $\omega_{x,j}\overline{\omega}_{x,j} = 0$ for each $j \in [\log n]$. Let us explain this in detail. The axioms are of the form $\bigvee_{i\in[\log n]} \omega_{j,i}^{1-a_i} \vee \nu_{j,a}$ which becomes $\omega_{j,1}^{1-a_1} + \cdots + \omega_{j,\log n}^{1-a_{\log n}} + \nu_{j,a} \geq 1$. We now lift through by $\omega_{j,1}^{a_1}$, $\ldots$, $\omega_{j,\log n}^{a_{\log n}}$ to obtain $\omega_{x,1}^{a_1} \ldots \omega_{x,\log n}^{a_{\log n}} \leq \omega_{x,1}^{a_1} \ldots \omega_{x,\log n}^{a_{\log n}} \nu_{x,a} \leq \nu_{x,a}$. □

The unary *Ordering Principle (*OP*) with equality* has the following set of SA axioms:

$$\begin{gathered}
\mathit{self} : v_{i,i} = 0 \quad \forall\, i \in n \\
\mathit{trans} : v_{i,k} - v_{i,j} - v_{j,k} + 1 \geq 0 \quad \forall\, i,j,k \in [n] \\
\mathit{impl} : v_{i,j} - w_{i,j} \geq 0 \quad \forall\, i,j \in [n] \\
\mathit{lower} : \textstyle\sum_{i\in[n]} w_{i,j} - 1 = 0 \quad \forall\, j \in [n]
\end{gathered}$$

Note that we need the w-variables since we use the equality form. Axioms of the form $\sum_{i\in[n]} x_{i,j} - 1 = 0$ made just from v-variables are plainly incompatible with, e.g., transitivity. Strictly speaking Sherali-Adams is defined for inequalities only. An equality axiom $a = 0$ is simulated by the two inequalities $a \geq 0, -a \geq 0$, which we refer to as the *positive* and *negative* instances of that axiom, respectively. Also, note that we have used $v_{i,j} + \overline{v}_{i,j} = 1$ to derive this formulation. We call two product terms *isomorphic* if one product term can be gotten from the other by relabelling the indices appearing in the subscripts by a permutation.

Theorem 4.3. *The SA rank of the* OP *with equality is at most* 2 *and SA size at most polynomial in* n*.*

Proof. Note that if the polytope $\mathcal{P}_2^{\mathsf{OP}}$ is nonempty there must exist a point where any isomorphic variables are given the same value. We can find such a point by averaging an asymmetric valuation over all permutations of $[n]$.
So suppose towards a contradiction there is such a symmetric point. First note $v_{i,i} = w_{i,i} = 0$ by *self* and *impl*. We start by lifting the jth instance of *lower* by $v_{i,j}$ to get

$$w_{i,j}v_{i,j} + \sum_{k\neq i,j} w_{k,j}v_{i,j} = v_{i,j}.$$

Equating (by symmetry with respect to k) the product terms $w_{k,j}v_{i,j}$ this is actually

$$w_{i,j}v_{i,j} + (n-2)w_{k,j}v_{i,j} = v_{i,j}.$$

Lift this by $w_{k,j}$ to get

$$w_{k,j}w_{i,j}v_{i,j} + (n-2)w_{k,j}v_{i,j} = w_{k,j}v_{i,j}.$$

We can delete the leftmost product term by proving it must be 0. Let us take an instance of *lower* lifted by $w_{k,j}v_{i,j}$ for any $k \neq i, j$ along with an instance of monotonicity $w_{k,j}w_{m,j}v_{i,j} \geq 0$ for every $m \neq j, k$:

$$\begin{aligned} w_{k,j}v_{i,j}\left(1 - \sum_{m\neq j} w_{m,j}\right) + \sum_{m\neq j,k,i} w_{k,j}w_{m,j}v_{i,j} \\ = -\sum_{m\neq k,j} w_{k,j}w_{m,j}v_{i,j} + \sum_{m\neq j,k,i} w_{k,j}w_{m,j}v_{i,j} \\ = -w_{k,j}w_{i,j}v_{i,j}. \end{aligned} \tag{4.1}$$

The left hand side of this equation is greater than 0 so we can deduce $w_{k,j}w_{i,j}v_{i,j} = 0$.

This results in

$$(n-2)w_{k,j}v_{i,j} = w_{k,j}v_{i,j} \quad \text{which is} \quad w_{k,j}v_{i,j} = 0.$$

We lift *impl* by $w_{i,j}$ to obtain $w_{i,j} \leq w_{i,j}v_{i,j}$. Monotonicity gives us the opposite inequality and we can proceed as if we had the equality $w_{k,j}v_{k,j} = w_{k,j}$ (as we are using equality as shorthand for inequality in both directions) .
So repeating the derivation of $w_{k,j}v_{i,j} = 0$ for every $i \neq k$ and then adding $w_{k,j}v_{k,j} = w_{k,j}$ gets us $\sum_m w_{k,j}v_{m,j} = w_{k,j}$. Repeating this again for every k and summing up gives

$$0 = \sum_{k,m} w_{k,j}v_{m,j} - \sum_k w_{k,j} = \sum_{k,m} w_{k,j}v_{m,j} - 1$$

with the last equality coming from the addition of the positive *lower* instance $\sum_k w_{k,j} - 1 = 0$. Finally adding the lifted *lower* instance $v_{m,j} - \sum_k w_{k,j}v_{m,j} = 0$ for every m gives

$$\sum_m v_{m,j} = 1. \tag{4.2}$$

By lifting the *trans* axiom $v_{i,k} - v_{i,j} - v_{j,k} + 1 \geq 0$ by $v_{j,k}$ we get

$$v_{i,k}v_{j,k} - v_{i,j}v_{j,k} \geq 0. \tag{4.3}$$

Now, due to a manipulation similar to Equation (4.1) using Equation (4.2)

$$\begin{aligned} v_{k,j}v_{i,j}\left(1 - \sum_{m\neq j} v_{m,j}\right) + \sum_{m\neq j,k,i} v_{k,j}v_{m,j}v_{i,j} \\ = -\sum_{m\neq k,j} v_{k,j}v_{m,j}v_{i,j} + \sum_{m\neq j,k,i} v_{k,j}v_{m,j}v_{i,j} \\ = -v_{k,j}v_{i,j}v_{i,j} \qquad (4.4) \\ = -v_{k,j}v_{i,j}. \qquad (4.5) \end{aligned}$$

Thus $v_{i,k}v_{j,k}$ must be zero whenever $i \neq j$. Along with Equation (4.3) we derive $v_{i,j}v_{j,k} = 0$. Noting $v_{i,j}v_{j,i} = 0$ follows from *trans* and *self*, we lift Equation (4.2) by $v_{j,x}$ for some x to get

$$v_{j,x}\sum_m v_{m,j} = \sum_{m\neq x,j} v_{m,j}v_{j,x} = v_{j,x}$$

where we know the left hand side is zero (Equation (4.3)). Thus we can derive $v_{i,j} = 0$ for any i and j, resulting in a contradiction when combined with Equation (4.2). □

Before we derive our corollary, let us explicitly give the SA axioms of BinOP_n.

$$\begin{gathered} \textit{self} : \nu_{i,i} = 0 \quad \forall\, i \in n \\ \textit{trans} : \nu_{i,k} - \nu_{i,j} - \nu_{j,k} + 1 \geq 0 \quad \forall\, i,j,k \in [n] \\ \textit{impl} : \sum_{i \in [\log n]} \omega_{i,j}^{1-a_i} + \nu_{j,a} \geq 0 \quad \forall\, j \in [n] \\ \text{where } a_1 \dots a_{\log n} = \mathrm{bin}(a) \end{gathered}$$

Corollary 4.3. *The binary encoding of the Ordering Principle,* BinOP_n*, has SA rank at most* $2 \log n$ *and SA size at most polynomial in* n*.*

Proof. Immediate from Lemma 4.7. □

4.5 SA+Squares

In this section we consider a proof system, SA+Squares, based on inequalities of multilinear polynomials. We now consider axioms as degree-1 polynomials in some set of variables and refutations as polynomials in those same variables. Then this system is gotten from SA by allowing addition of (linearised) squares of polynomials. In terms of strength this system will be strictly stronger than SA and at most as strong as Lasserre (also known as Sum-of-Squares), although we do not at this point see an exponential separation between SA+Squares and Lasserre. See [5, 57] for more on the Lasserre proof system and [59] for tight degree lower bound results.

Consider the polynomial $w_{i,j}v_{i,j} - w_{i,j}v_{i,k}$. The square of this is

$$w_{i,j}v_{i,j}w_{i,j}v_{i,j} + w_{i,j}v_{i,k}w_{i,j}v_{i,k} - 2w_{i,j}v_{i,j}w_{i,j}v_{i,k}.$$

Using idempotence this linearises to $w_{i,j}v_{i,j} + w_{i,j}v_{i,k} - 2w_{i,j}v_{i,j}v_{i,k}$, and so we know that this last polynomial is non-negative for all 0/1 settings of the variables.
A *degree-d* SA+Squares refutation of a set of linear inequalities (over terms) $q_1 \geq 0, \dots, q_x \geq 0$ is an equation of the form

$$\sum_{i=1}^{x} p_i q_i + \sum_{i=1}^{y} r_i^2 = -1 \tag{4.6}$$

where the p_i are polynomials with nonnegative coefficients and the degree of the polynomials $p_i q_i, r_i^2$ is at most d. We want to underline that we now consider a (product) term like $w_{i,j}v_{i,j}v_{i,k}$ as a product of its constituent variables, that is genuinely a term in the sense of part of a polynomial. This is opposed to the preceding sections in which we

viewed it as a single variable $Z_{w_{i,j} \wedge v_{i,j} \wedge v_{i,k}}$. The translation from the degree discussed here to SA rank previously introduced may be paraphrased by "rank = degree − 1".

We note that the unary PHP_n^{n+1} becomes easy in this stronger proof system (see, e.g., Example 2.1 in [44]) while we shall see that the LOP remains hard (in terms of degree). The following is based on Example 2.1 in [44].

Theorem 4.4. *The* BinPHP_n^{n+1} *has an SA* +Squares *refutation of degree* $2\log n + 1$ *and size* $O(n^3)$.

Proof. For short let $m = n+1$ denote the number of pigeons. We begin by squaring the polynomial

$$1 - \sum_{i=1}^{m} \prod_{j=1}^{\log n} \omega_{i,j}^{a_j}$$

to get the degree $2\log n$, size quadratic in m inequality

$$1 - 2\sum_{i=1}^{m} \prod_{j=1}^{\log n} \omega_{i,j}^{a_j} + \sum_{1 \le i,i' \le m} \left(\prod_{j=1}^{\log n} \omega_{i,j}^{a_j} \right) \left(\prod_{j=1}^{\log n} \omega_{i',j}^{a_j} \right) \ge 0 \tag{4.7}$$

for every hole $a \in [n]$. On the other hand, by lifting each axiom

$$\sum_{j=1}^{\log n} \omega_{i,j}^{1-a_j} + \sum_{j=1}^{\log n} \omega_{i',j}^{1-a_j} \ge 1 \qquad \text{(whenever } i \ne i'\text{)}$$

by $\left(\prod_{j=1}^{\log n} \omega_{i,j}^{a_j}\right) \left(\prod_{j=1}^{\log n} \omega_{i',j}^{a_j}\right)$ we find $0 \ge \left(\prod_{j=1}^{\log n} \omega_{i,j}^{a_j}\right) \left(\prod_{j=1}^{\log n} \omega_{i',j}^{a_j}\right)$, in degree $2\log n + 1$. Adding these inequalities to (4.7) gives

$$1 - \sum_{i=1}^{m} \prod_{j=1}^{\log n} \omega_{i,j}^{a_j} \ge 0$$

in size again quadratic in m. Iterating this for every hole $a \in [n]$ we find

$$n - \sum_{a=1}^{n} \sum_{i=1}^{m} \prod_{j=1}^{\log n} \omega_{i,j}^{a_j} \ge 0 \tag{4.8}$$

in cubic size.

Note that for any pigeon $i \in [m]$, we can find in SA the linearly sized equality

$$\sum_{a=1}^{n} \prod_{j=1}^{\log n} \omega_{i,j}^{a_j} = 1. \tag{4.9}$$

in size linear in n.

This is done by induction on the number of bits involved (the range of j in the summation). For the base case of just $j = 1$ we clearly have

$$\omega_{i,1} + (1 - \omega_{i,1}) = 1.$$

Now suppose that for $k < \log n$, we have $\sum_{a\in[2^k]} \prod_{j=1}^{k} \omega_{i,j}^{a_j} = 1$. Multiplying both sides by $1 = \omega_{i,(k+1)} + (1 - \omega_{i,(k+1)})$ gets the inductive step. The final term is of size $O(2^{\log n}) = O(n)$.

Summing 4.9 for every such hole i we find

$$\sum_{i=1}^{m} \sum_{a=1}^{n} \prod_{j=1}^{\log n} \omega_{i,j}^{a_j} \geq m. \tag{4.10}$$

Adding 4.10 to 4.8, we get the desired contradiction, $n - m \geq 0$. □

This last theorem, combined with the exponential SA size lower bound given in Theorem 4.1, shows us that SA+Squares is exponentially separated from SA (in terms of size).

We now turn our attention to LOP, whose SA axioms we reproduce to refresh the reader's memory.

$$\begin{gathered}
\textit{self} : v_{i,i} = 0 \quad \forall\, i \in n \\
\textit{trans} : v_{i,k} - v_{i,j} - v_{j,k} + 1 \geq 0 \quad \forall\, i,j,k \in [n] \\
\textit{impl} : v_{i,j} - w_{i,j} \geq 0 \quad \forall\, i,j \in [n] \\
\textit{total} : v_{i,j} + v_{j,i} - 1 \geq 0 \quad \forall\, i \neq j \in [n] \\
\textit{lower} : \sum_{i\in[n]} w_{i,j} - 1 \geq 0 \quad \forall\, j \in [n]
\end{gathered}$$

We give our lower bound for the unary LOP by producing a linear function ν (which we will call a *valuation*) from terms into $\mathbb{R}$ such that

1. for each axiom $p \geq 0$ and every term X with $deg(Xp) \leq d$ we have $\nu(Xp) \geq 0$, and
2. we have $\nu(r^2) \geq 0$ whenever $deg(r^2) \leq d$.
3. $\nu(1) = 1$.

The existence of such a valuation clearly implies that a degree-d SA+Squares refutation cannot exist, as it would result in a contradiction when applied to both sides of eq. (4.6).

To verify that $\nu(r^2) \geq 0$ whenever $deg(r^2) \leq d$ we show that the so-called *moment-matrix* $\mathcal{M}_\nu$ is positive semidefinite. The degree-d moment matrix is defined to be the

symmetric square matrix whose rows and columns are indexed by terms of size at most $d/2$ and each entry is the valuation of the product of the two terms indexing that entry. Given any polynomial σ of degree at most $d/2$ let c be its vector of coefficients. Then if $\mathcal{M}_v$ is positive semidefinite:

$$\nu(\sigma^2) = \sum_{deg(T_1),deg(T_2)\leq d/2} c(T_1)c(T_2)v(T_1T_2) = c^\top \mathcal{M}_v c \geq 0.$$

(For more on this see e.g. [57], Section 2.)

Theorem 4.5. *There is no SA* +Squares *refutation of the (unary)* LOP *with degree at most* $(n-3)/2$.

Proof. For each term T, let $\nu(T)$ be the probability that T is consistent with a permutation on the n elements taken uniformly at random or, in other words, the number of permutations consistent with T divided by $n!$. Here we view $w_{x,y}$ as equal to $v_{x,y}$. This valuation trivially satisfies the lifts of the *self*, *trans* and *total* axioms as they are satisfied by each permutation (linear order). It satisfies the lifts of the *impl* axioms by construction. We now claim that the lifts of the *lower* axioms (those containing only w variables) of degree up to $\frac{n-3}{2}$ are also satisfied by $v(.)$. Indeed, let us consider the lifting by T of the *lower* axiom for x

$$\sum_{y=1}^{n} Tw_{x,y} \geq T. \tag{4.11}$$

Since T mentions at most $n-3$ elements, there must be at least two $y_1 \neq y_2$ that are different from all of them and from x. For any permutation that is consistent with T, the probability that each of the y_1 and y_2 is smaller than x is precisely a half, and thus

$$\nu(Tw_{x,y_1}) + \nu(Tw_{x,y_2}) = \nu(T).$$

Therefore the valuation of the LHS of (4.11) is always greater than or equal to the valuation of T.

Finally, we need to show that the valuation is consistent with the non-negativity of (the linearisation of) any squared polynomial. It is easy to see that the moment matrix for ν can be written as

$$\frac{1}{n!}\sum_{\sigma} V_\sigma V_\sigma^T$$

where the summation is over all permutations on n elements and for a permutation σ, V_σ is its characteristic vector. The characteristic vector of a permutation σ is a Boolean

column vector indexed by terms and whose entries are 1 or 0 depending on whether the respective index term is consistent or not with the permutation σ. Clearly the moment matrix is positive semidefinite being a sum of (rank one) positive semidefinite matrices. □

An alternative formulation of the Least Ordering Principle asks that the order be total, and this is enforced with axioms *anti-sym* of the form $P_{i,j} \vee P_{j,i}$, or $P_{i,j} + P_{j,i} \geq 1$, for $i \neq j \in [n]$. Let us call this alternative formulation OP. Ideally, lower bounds should be proved for OP, because they are potentially stronger. Conversely, upper bounds are stronger when they are proved on the ordinary OP, without the total order. Looking into the last proof, one sees that the lifts of *anti-sym* are satisfied as we derive our valuation exclusively from total orders. This is interesting because an upper bound in Sum of Squares of order $\sqrt{n}\log(n)$ is known for OP_n [71]. It is proved for a different formulation of OP_n than ours, which we relate here, and purposefully give in the variables $x_{i,j}$ (as well as the continuous z_j) instead of our $P_{i,j}$.

$$\begin{aligned} x_{i,j} + x_{j,i} &= 1 \qquad \text{for all distinct } i,j \in [n] \\ x_{i,j}x_{j,k}(1 - x_{i,k}) &= 0 \qquad \text{for all distinct } i,j,k \in [n] \\ \sum_{i\in[n],i\neq i} x_{i,j} &= 1 + z_j^2 \end{aligned}$$

We show that the difference is only superficial. Note that anything we can prove using transitivity of the form $x_{i,j}x_{j,k}(1 - x_{i,k}) = 0$ we can prove using $P_{i,k} - P_{i,j} - P_{j,k} \geq -1$. That $P_{i,j}P_{j,k} \geq P_{i,j}P_{j,k}P_{i,k}$ comes from monotonicity, and the opposite inequality comes from lifting by $P_{i,j}P_{j,k}$:

$$-P_{i,j}P_{j,k} \leq P_{i,j}P_{j,k}P_{i,k} - 2P_{i,j}P_{j,k} \implies P_{i,j}P_{j,k} \leq P_{i,j}P_{j,k}P_{i,k}.$$

Potechin's proof moves along the following lines. Define an operator E on terms that behaves the same as the v used in Theorem 4.5, but

1. If some z_j appears with degree 1 in T, then $E[T] = 0$, and
2. If T is of the form $z_j^2 T'$ for some j and T', $E[T] = E\left[\left(\sum_{i\in[n],i\neq i} x_{ij} - 1\right) T'\right]$

Potechin proves the following.

Lemma 4.8 (Lemma 4.2 in [71]). *There exists a polynomial g, only in the variables $x_{i,j}$ and of degree $O(\sqrt{n}\log n)$ such that*

$$E\left[\left(\sum_{i\neq j} x_{i,j} - 1\right) g^2\right] = v\left(\left(\sum_{i\neq j} x_{i,j} - 1\right) g^2\right) < 0.$$

Potechin then proves the following SOS identity using only the totality and transitivity axioms (which exist also in our formulation). Note S_k is the symmetric group on the elements of $[k]$.

Lemma 4.9 (Lemma 4.7 in [71]). *For all $A = \{i_1, i_2, \ldots, i_k\} \subseteq [n]$, there exists a degree $k+2$ proof that*

$$\sum_{\pi \in S_k} \prod_{j=1}^{k-1} x_{i_{\pi(j)} i_{\pi(j+1)}} = 1.$$

Finally, Potechin proves that the 'symmetric group average' of a polynomial can be shown to be equal to its valuation.

Lemma 4.10 (Lemma 4.8 in [71]). *For any polynomial p of degree d in the variables x_{ij}, there exists a proof of at most degree $3d+2$ that*

$$\frac{1}{n!} \sum_{\pi \in S_n} \pi(p) = v(p)$$

(where the action of S_n is to permute the indices in the monomials of p).

Lemma 4.8 and 4.10 together furnish a SOS refutation of the required form. Thus, Theorem 4.5, together with [71], shows a quadratic rank separation between SA+Squares and Sum of Squares.

4.6 Conclusions

In this section, we showed that, for SA, the movement between unary and binary encodings is not monotonic in terms of complexity - we saw that the PHP became exponentially harder in terms of size, and that the LOP became exponentially easier in terms of rank. This leaves one wondering if there could be any completely systematic categorisation of the complexity of the unary and binary encodings at all - the results of this chapter suggest, but do not prove, that there can be no such thing for SA.

The results of this vein in this chapter, however, are given for SA and its strengthenings only, and we are less sure about the analogous situation in different proof systems. This brings us to our first open problem of the epilogue of this thesis.

Chapter 5

Further directions

In the previous Chapter 4, we showed that the relationship between binary and unary encodings is a function of the principle at hand. In particular, we showed that for the LOP (in SA) the unary version is harder in terms of degree than its binary relative. However, we do not know if this is possible in Resolution.

Open Problem 1. *What is the relationship, if any, between the binary and unary encodings of a principle in Resolution? Are there any examples of a principle where the unary encoding requires higher width than the binary encoding?*

In the same Chapter, we pointed out a system - namely, SA+Squares - that lies somewhere in between SA and SOS in terms of power. We then provided an exponential separation between SA+Squares and SA across both rank and size. We also noted a quadratic degree separation between SA+Squares and SOS. But we are not sure if there are stronger examples of a separation in this direction.

Open Problem 2. *Is there a strong degree or size separation of SA+Squares from SOS?*

In Section 3.2 we began to lay out a framework to prove lower bounds against SOS for a large class of principles. As noticed in the conclusions to that section, the completion of that result is in itself an open problem. However, in order to achieve this completion, you would have to along the way give some new model-theoretic distinction between two of the most important principles in Proof Complexity.

Open Problem 3. *What exactly is the model-theoretic difference between the* PHP *and* LOP *that causes the latter to be harder for SOS than the former? Also, is there a model-theoretic explanation for their different complexities in SA, in terms of the movement between unary and binary encodings?*

In Chapter 2 we provided some tight lower bounds - for example, in the case of the SPHP, we demonstrated an $O(\log n)$ depth upper bound and a matching $\Omega(\log(n))$ lower bound. However, for Tseitin, there is a logarithmic gap between, for example, the $\Omega(\log(n))$ lower bound given in Corollary 2.4 and the $O(\log^2(n))$ upper bound given (for constant degree graphs) in the seminal [7]. This leads us to question:

Open Problem 4. *Where, exactly, does the SP depth of* $\mathsf{Ts}(G, \omega)$ *lie in the interval* $[\log(|G|),\ \log(|G|)^2]$*?*

Given how much power our techniques seem to leave on the table (as discussed in the conclusion to Chapter 2), we conjecture it should lie at the high end of this interval.

In Section 1.2.3, we began by defining the *Positivstellensatz* proof system, and then defining the SOS proof system as a restricted version. The restriction seems to be quite severe - in the case of the full Positivstellensatz we are allowed to use, if promised the nonnegativity of some polynomials $q_1 \geq 0, \ldots, q_m \geq 0$, that $\prod_{i \in I} q_i \geq 0$ also, for any index set $I \subseteq [m]$. In the case of SOS, which has certainly received the most attention, we are only allowed to use only the nonnegativity that was given. It is interesting to note that the reality of this restriction has not been properly demonstrated:

Open Problem 5. *Is there a principle admitting more efficient refutations by Positivstellensatz than by SOS? Or is it the case that SOS can actually simulate Positivstellensatz?*

We finish by returning to the dichotomy given in [28], where it is shown that for SA, a principle is hard if and only if it has only infinite models. Systems like SOS and SA+Squares cannot exhibit the same complexity gap, as both have a constant degree upper bound for the PHP, which has only infinite models. However, to our knowledge, CP might have the same gap.

Open Problem 6. *Let* $\mathcal{T}$ *be a theory having only infinite models. Does* $\mathcal{T}_n$ *have non-constant CP rank?*

Bibliography

[1] Alon, N., and Füredi, Z. Covering the cube by affine hyperplanes. *Eur. J. Comb. 14*, 2 (1993), 79–83.

[2] Atserias, A., Lauria, M., and Nordström, J. Narrow proofs may be maximally long. *ACM Trans. Comput. Log. 17*, 3 (2016), 19:1–19:30.

[3] Banks, J., Kleinberg, R., and Moore, C. The lov\'asz theta function for random regular graphs and community detection in the hard regime. *arXiv preprint arXiv:1705.01194* (2017).

[4] Barak, B., Hopkins, S. B., Kelner, J. A., Kothari, P., Moitra, A., and Potechin, A. A nearly tight sum-of-squares lower bound for the planted clique problem. *CoRR abs/1604.03084* (2016).

[5] Barak, B., and Steurer, D. Sum-of-squares proofs and the quest toward optimal algorithms. In *Proceedings of International Congress of Mathematicians (ICM)* (2014), vol. IV, pp. 509–533.

[6] Barak, B., and Steurer, D. The unique games and sum of squares: A love hate relationship, 2016.

[7] Beame, P., Fleming, N., Impagliazzo, R., Kolokolova, A., Pankratov, D., Pitassi, T., and Robere, R. Stabbing planes. In *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA* (2018), A. R. Karlin, Ed., vol. 94 of *LIPIcs*, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, pp. 10:1–10:20.

[8] Beame, P., and Pitassi, T. Propositional proof complexity: Past, present, and future. 42–70.

[9] Ben-sasson, E., and Wigderson, A. Short proofs are narrow - resolution made simple. In *Journal of the ACM* (1999), pp. 517–526.

[10] Berkholz, C. The relation between polynomial calculus, sherali-adams, and sum-of-squares proofs. In *35th Symposium on Theoretical Aspects of Computer Science (STACS 2018)* (2018), Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[11] Blake, A. *Canonical expressions in Boolean algebra*. PhD thesis, The University of Chicago, 1937.

[12] Bonacina, I., and Galesi, N. A framework for space complexity in algebraic proof systems. *J. ACM 62*, 3 (2015), 23:1–23:20.

[13] Bonacina, I., Galesi, N., and Thapen, N. Total space in resolution. *SIAM J. Comput. 45*, 5 (2016), 1894–1909.

[14] Bonet, M. L., and Galesi, N. A study of proof search algorithms for resolution and polynomial calculus. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)* (1999), IEEE, pp. 422–431.

[15] Buresh-Oppenheim, J., Galesi, N., Hoory, S., Magen, A., and Pitassi, T. Rank bounds and integrality gaps for cutting planes procedures. *Theory of Computing 2*, 4 (2006), 65–90.

[16] Buss, S. R., and Pitassi, T. Resolution and the weak pigeonhole principle. In *Computer Science Logic, 11th International Workshop, CSL '97, Annual Conference of the EACSL, Aarhus, Denmark, August 23-29, 1997, Selected Papers* (1997), pp. 149–156.

[17] Carroll, T., Cooper, J., and Tetali, P. Counting antichains and linear extensions in generalizations of the boolean lattice, 2009.

[18] Chvátal, V. Edmonds polytopes and a hierarchy of combinatorial problems. *Discrete mathematics 4*, 4 (1973), 305–337.

[19] Chvátal, V., Cook, W., and Hartmann, M. On cutting-plane proofs in combinatorial optimization. *Linear algebra and its applications 114* (1989), 455–499.

[20] Cook, S. A., and Reckhow, R. A. The relative efficiency of propositional proof systems. *The Journal of Symbolic Logic 44*, 1 (1979), 36–50.

[21] Cook, W., Coullard, C. R., and Turán, G. On the complexity of cutting-plane proofs. *Discrete Appl. Math. 18*, 1 (1987), 25–38.

[22] Dadush, D., and Tiwari, S. On the complexity of branching proofs. In *35th Computational Complexity Conference, CCC 2020, July 28-31, 2020, Saarbrücken, Germany (Virtual Conference)* (2020), S. Saraf, Ed., vol. 169 of *LIPIcs*, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, pp. 34:1–34:35.

[23] Dantchev, S., and Martin, B. Rank complexity gap for lovász-schrijver and sherali-adams proof systems. *computational complexity 22*, 1 (2013), 191–213.

[24] Dantchev, S. S. Rank complexity gap for Lovász-Schrijver and Sherali-Adams proof systems. In *STOC '07: Proceedings of the thirty-ninth annual ACM symposium on Theory of computing* (New York, NY, USA, 2007), ACM Press, pp. 311–317.

[25] Dantchev, S. S., Galesi, N., Ghani, A., and Martin, B. Depth lower bounds in stabbing planes for combinatorial principles. In *39th International Symposium on Theoretical Aspects of Computer Science, STACS 2022, March 15-18, 2022, Marseille, France (Virtual Conference)* (2022), P. Berenbrink and B. Monmege, Eds., vol. 219 of *LIPIcs*, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, pp. 24:1–24:18.

[26] Dantchev, S. S., Galesi, N., and Martin, B. Resolution and the binary encoding of combinatorial principles. In *34th Computational Complexity Conference, CCC 2019, July 18-20, 2019, New Brunswick, NJ, USA.* (2019), pp. 6:1–6:25. See http://arxiv.org/abs/1809.02843.

[27] Dantchev, S. S., Ghani, A., and Martin, B. Sherali-adams and the binary encoding of combinatorial principles. In *LATIN 2020: Theoretical Informatics - 14th Latin American Symposium, São Paulo, Brazil, January 5-8, 2021, Proceedings* (2020), Y. Kohayakawa and F. K. Miyazawa, Eds., vol. 12118 of *Lecture Notes in Computer Science*, Springer, pp. 336–347.

[28] Dantchev, S. S., and Martin, B. Rank complexity gap for lovász-schrijver and sherali-adams proof systems. *Computational Complexity 22*, 1 (2013), 191–213.

[29] Dantchev, S. S., Martin, B., and Rhodes, M. N. C. Tight rank lower bounds for the sherali-adams proof system. *Theor. Comput. Sci. 410*, 21-23 (2009), 2054–2063.

[30] Davis, M., Logemann, G., and Loveland, D. W. A machine program for theorem-proving. *Commun. ACM 5*, 7 (1962), 394–397.

[31] Davis, M., and Putnam, H. A computing procedure for quantification theory. *J. ACM 7*, 3 (1960), 201–215.

[32] Dunkel, J. *The Gomory-Chvátal closure: Polyhedrality, complexity, and extensions.* PhD thesis, Massachusetts Institute of Technology, 2011.

[33] Eisenbrand, F., and Schulz, A. S. Bounds on the chvátal rank of polytopes in the 0/1-cube. *Combinatorica 23*, 2 (2003), 245–261.

[34] Filmus, Y., Lauria, M., Nordström, J., Ron-Zewi, N., and Thapen, N. Space complexity in polynomial calculus. *SIAM J. Comput. 44*, 4 (2015), 1119–1153.

[35] Fleming, N., Göös, M., Impagliazzo, R., Pitassi, T., Robere, R., Tan, L., and Wigderson, A. On the power and limitations of branch and cut. In *36th Computational Complexity Conference, CCC 2021, July 20-23, 2021, Toronto, Ontario, Canada (Virtual Conference)* (2021), V. Kabanets, Ed., vol. 200 of *LIPIcs*, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, pp. 6:1–6:30.

[36] Fleming, N., and Pitassi, T. Reflections on proof complexity and counting principles. *Alasdair Urquhart on Nonclassical and Algebraic Logic and Complexity of Proofs* (2022), 497–520.

[37] Galesi, N., Pudlák, P., and Thapen, N. The space complexity of cutting planes refutations. In *30th Conference on Computational Complexity, CCC 2015, June 17-19, 2015, Portland, Oregon, USA* (2015), D. Zuckerman, Ed., vol. 33 of *LIPIcs*, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, pp. 433–447.

[38] Galil, Z. The complexity of resolution procedures for theorem proving in the propositional calculus. Tech. rep., Cornell University, 1975.

[39] Gomory, R. E. Solving linear programming problems in integers. In *Combinatorial Analysis, Proceedings of Symposia in Applied Mathematics* (Providence, RI, 1960), R. Bellman and M. Hall, Eds., vol. 10.

[40] Gomory, R. E. An algorithm for integer solutions to linear programs. In *Recent advances in mathematical programming*. McGraw-Hill, New York, 1963, pp. 269–302.

[41] Grigoriev, D. Complexity of positivstellensatz proofs for the knapsack. *computational complexity 10*, 2 (Dec 2001), 139–154.

[42] Grigoriev, D. Linear lower bound on degrees of positivstellensatz calculus proofs for the parity. *Theor. Comput. Sci. 259*, 1-2 (2001), 613–622.

[43] Grigoriev, D., Hirsch, E. A., and Pasechnik, D. V. Complexity of semi-algebraic proofs. In *Annual Symposium on Theoretical Aspects of Computer Science* (2002), Springer, pp. 419–430.

[44] Grigoriev, D., Hirsch, E. A., and Pasechnik, D. V. Complexity of semi-algebraic proofs. In *STACS '02: Proceedings of the 19th Annual Symposium on Theoretical Aspects of Computer Science* (London, UK, 2002), Springer-Verlag, pp. 419–430.

[45] Haken, A. The intractability of resolution. *Theor. Comput. Sci. 39* (1985), 297–308.

[46] Hall, P. On Representatives of Subsets. *Journal of the London Mathematical Society s1-10*, 1 (01 1935), 26–30.

[47] Hirsch, E. Optimal acceptors and optimal proof systems. *Theory and Applications of Models of Computation* (2010), 28–39.

[48] Hrubes, P., and Pudlák, P. Random formulas, monotone circuits, and interpolation. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017* (2017), C. Umans, Ed., IEEE Computer Society, pp. 121–131.

[49] Impagliazzo, R., Pitassi, T., and Urquhart, A. Upper and lower bounds for tree-like cutting planes proofs. In *Proceedings Ninth Annual IEEE Symposium on Logic in Computer Science* (1994), IEEE, pp. 220–228.

[50] Josz, C., and Henrion, D. Strong duality in lasserre's hierarchy for polynomial optimization. *Optimization Letters 10*, 1 (Jan 2016), 3–10.

[51] Kautz, H. A., and Selman, B. Ten challenges redux: Recent progress in propositional reasoning and search. In *Principles and Practice of Constraint Programming - CP 2003, 9th International Conference, CP 2003, Kinsale, Ireland, September 29 - October 3, 2003, Proceedings* (2003), F. Rossi, Ed., vol. 2833 of *Lecture Notes in Computer Science*, Springer, pp. 1–18.

[52] Kavitha, T., Liebchen, C., Mehlhorn, K., Michail, D., Rizzi, R., Ueckerdt, T., and Zweig, K. A. Cycle bases in graphs characterization, algorithms, complexity, and applications. *Computer Science Review 3*, 4 (2009), 199–243.

[53] Kojevnikov, A. Improved lower bounds for tree-like resolution over linear inequalities. In *Theory and Applications of Satisfiability Testing - SAT 2007, 10th International Conference, Lisbon, Portugal, May 28-31, 2007, Proceedings* (2007), J. Marques-Silva and K. A. Sakallah, Eds., vol. 4501 of *Lecture Notes in Computer Science*, Springer, pp. 70–79.

[54] Krajícek, J. Discretely ordered modules as a first-order extension of the cutting planes proof system. *J. Symb. Log. 63*, 4 (1998), 1582–1596.

[55] Krajícek, J. Interpolation by a game. *Math. Log. Q. 44* (1998), 450–458.

[56] Krajíček, J., and Pudlák, P. Propositional proof systems, the consistency of first order theories and the complexity of computations. *The Journal of Symbolic Logic 54*, 3 (1989), 1063–1079.

[57] Lasserre, J. B. Global optimization with polynomials and the problem of moments. *SIAM Journal on optimization 11*, 3 (2001), 796–817.

[58] Laurent, M. A comparison of the sherali-adams, lovász-schrijver, and lasserre relaxations for 0-1 programming. *Mathematics of Operations Research 28*, 3 (2003), 470–496.

[59] Lauria, M., and Nordström, J. Tight size-degree bounds for sums-of-squares proofs. *computational complexity 26*, 4 (Dec 2017), 911–948.

[60] Lauria, M., Pudlák, P., Rödl, V., and Thapen, N. The complexity of proving that a graph is ramsey. *Combinatorica 37*, 2 (2017), 253–268.

[61] Lee, J. R., Raghavendra, P., and Steurer, D. Lower bounds on the size of semidefinite programming relaxations. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing* (New York, NY, USA, 2015), STOC '15, Association for Computing Machinery, p. 567–576.

[62] Linial, N., and Radhakrishnan, J. Essential covers of the cube by hyperplanes. *Journal of Combinatorial Theory, Series A 109*, 2 (2005), 331–338.

[63] Lovász, L. On the shannon capacity of a graph. *IEEE Transactions on Information theory 25*, 1 (1979), 1–7.

[64] Lovász, L. *Large networks and graph limits*, vol. 60. American Mathematical Soc., 2012.

[65] Ma, T., and Wigderson, A. Sum-of-squares lower bounds for sparse PCA. *CoRR abs/1507.06370* (2015).

[66] Mattner, L., and Roos, B. Maximal probabilities of convolution powers of discrete uniform distributions. *Statistics & probability letters 78*, 17 (2008), 2992–2996.

[67] Mitzenmacher, M., and Upfal, E. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

[68] Mohanty, S., Raghavendra, P., and Xu, J. Lifting sum-of-squares lower bounds: degree-2 to degree-4. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing* (2020), pp. 840–853.

[69] Part, F., and Tzameret, I. Resolution with counting: Dag-like lower bounds and different moduli. *Comput. Complex. 30*, 1 (2021), 2.

[70] Potechin, A. Sum of squares lower bounds from symmetry and a good story. *CoRR abs/1711.11469* (2017).

[71] Potechin, A. Sum of squares bounds for the ordering principle. In *Proceedings of the 35th Computational Complexity Conference* (Dagstuhl, DEU, 2020), CCC '20, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[72] Pudlák, P. Proofs as games. *American Mathematical Monthly* (June-July 2000), 541–550.

[73] Raymond, A., Saunderson, J., Singh, M., and Thomas, R. R. Symmetric sums of squares over k-subset hypercubes. *Mathematical Programming 167*, 2 (2018), 315–354.

[74] Razborov, A. A. Flag algebras. *The Journal of Symbolic Logic 72*, 4 (2007), 1239–1282.

[75] Rhodes, M. Rank lower bounds for the Sherali-Adams operator. In *CiE* (2007), S. B. Cooper, B. Löwe, and A. Sorbi, Eds., vol. 4497 of *Lecture Notes in Computer Science*, Springer, pp. 648–659.

[76] Rhodes, M. N. C. On the chvátal rank of the pigeonhole principle. *Theor. Comput. Sci. 410*, 27-29 (2009), 2774–2778.

[77] Riis, S. A complexity gap for tree resolution. *Computational Complexity 10*, 3 (2001), 179–209.

[78] Riis, S. On the asymptotic nullstellensatz and polynomial calculus proof complexity. In *Logic in Computer Science, 2008. LICS'08. 23rd Annual IEEE Symposium on* (2008), IEEE, pp. 272–283.

[79] Robinson, J. A. A machine-oriented logic based on the resolution principle. *J. ACM 12*, 1 (1965), 23–41.

[80] Rothvoss, T. The lasserre hierarchy in approximation algorithms.

[81] Schrijver, A., et al. On cutting planes. *Combinatorics 79* (1980), 291–296.

[82] Selman, B., Kautz, H. A., and McAllester, D. A. Ten challenges in propositional reasoning and search. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, IJCAI 97, Nagoya, Japan, August 23-29, 1997, 2 Volumes* (1997), Morgan Kaufmann, pp. 50–54.

[83] Sherali, H. D., and Adams, W. P. A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. *SIAM J. Discrete Math. 3*, 3 (1990), 411–430.

[84] Tseitin, G. S. On the complexity of proof in prepositional calculus. *Zapiski Nauchnykh Seminarov POMI 8* (1968), 234–259.

[85] Urquhart, A. Hard examples for resolution. *J. ACM 34*, 1 (1987), 209–219.

[86] Van Lint, J., and Wilson, R. A course in combinatorics. *Cambridge Univ. Press, New York* (1992).

[87] van Lint, J. H., and Wilson, R. M. *A Course in Combinatorics*. Cambridge University Press, Cambridge, U.K.; New York, 2001.

[88] Yehuda, G., and Yehudayoff, A. A lower bound for essential covers of the cube. *arXiv preprint arXiv:2105.13615* (2021).