Optimal sparsification for some binary CSPs using low-degree polynomials

Optimal Sparsification for Some Binary CSPs Using Low-degree Polynomials*

Bart M. P. Jansen[1] **and Astrid Pieterse**[2]

1 **Eindhoven University of Technology**
P.O. Box 513, Eindhoven, The Netherlands
b.m.p.jansen@tue.nl

2 **Eindhoven University of Technology**
P.O. Box 513, Eindhoven, The Netherlands
a.pieterse@tue.nl

Abstract

This paper analyzes to what extent it is possible to efficiently reduce the number of clauses in NP-hard satisfiability problems, without changing the answer. Upper and lower bounds are established using the concept of kernelization. Existing results show that if NP $\not\subseteq$ coNP/poly, no efficient preprocessing algorithm can reduce n-variable instances of CNF-SAT with d literals per clause, to equivalent instances with $\mathcal{O}(n^{d-\varepsilon})$ bits for any $\varepsilon > 0$. For the Not-All-Equal SAT problem, a compression to size $\widetilde{\mathcal{O}}(n^{d-1})$ exists. We put these results in a common framework by analyzing the compressibility of binary CSPs. We characterize constraint types based on the minimum degree of multivariate polynomials whose roots correspond to the satisfying assignments, obtaining (nearly) matching upper and lower bounds in several settings. Our lower bounds show that not just the number of constraints, but also the encoding size of individual constraints plays an important role. For example, for Exact Satisfiability with unbounded clause length it is possible to efficiently reduce the number of constraints to $n+1$, yet no polynomial-time algorithm can reduce to an equivalent instance with $\mathcal{O}(n^{2-\varepsilon})$ bits for any $\varepsilon > 0$, unless NP $\subseteq$ coNP/poly.

1998 ACM Subject Classification F.2.2 Nonnumerical Algorithms and Problems, F.4.1 Mathematical Logic

Keywords and phrases constraint satisfaction problem, sparsification, satisfiability, kernelization

1 Introduction

The goal of sparsification is to make an object such as a graph or logical structure less dense, without changing the outcome of a computational task of interest. Sparsification can be used to speed up the solution of NP-hard problems, by sparsifying a problem instance before solving it. The notion of kernelization, originating in the field of parameterized complexity [9, 13, 14], facilitates a rigorous study of polynomial-time preprocessing for NP-hard problems and can be used to reason about (the impossibility of) sparsification. Over the last few years, our understanding of the power of polynomial-time data reduction has increased tremendously, as documented in recent surveys [5, 17, 23, 26]. By studying the kernelization complexity of a graph problem parameterized by the number of vertices, or of a logic problem parameterized by the number of variables, we can analyze its potential for sparsification.

* This work was supported by NWO Veni grant "Frontiers in Parameterized Preprocessing" and NWO Gravitation grant "Networks". This work was done in part while the first author was visiting the Simons Institute for the Theory of Computing.

The vast majority of the currently known results in this direction are negative [11, 19, 20, 21], stating that no nontrivial sparsification is possible under plausible complexity-theoretic assumptions. For example, Dell and van Melkebeek [11] obtained such a result for CNF-Satisfiability with clauses of size at most d (d-CNF-SAT), for each fixed $d \geq 3$. Assuming $\mathsf{NP} \not\subseteq \mathsf{coNP/poly}$, there is no polynomial-time algorithm that compresses any n-variable instance of d-CNF-SAT to an equivalent instance with $\mathcal{O}(n^{d-\varepsilon})$ bits for $\varepsilon > 0$. Since there are $\mathcal{O}(n^d)$ possible clauses of size at most d over n variables, the trivial compression scheme that outputs a bitstring of length $\mathcal{O}(n^d)$, denoting for each possible clause whether it occurs in the instance or not, is optimal up to $n^{o(1)}$ factors.

A problem for which nontrivial polynomial-time sparsification *is* possible was recently discovered by the current authors [21]. Any n-variable instance of the Not-All-Equal CNF-Satisfiability problem with clauses of size at most d (d-NAE-SAT) can efficiently be compressed to an equivalent instance with $\mathcal{O}(n^{d-1})$ clauses, which can be encoded in $\mathcal{O}(n^{d-1} \log n)$ bits. The preprocessing algorithm is based on a linear-algebraic lemma by Lovász [27] to identify clauses that are implied by others, allowing a reduction from $\Theta(n^d)$ clauses to $\mathcal{O}(n^{d-1})$. This sparsification for d-NAE-SAT forms the starting point for this work. Since d-CNF-SAT and d-NAE-SAT can both be seen as constraint satisfaction problems (CSPs) with a binary domain, it is natural to ask whether the positive results for d-NAE-SAT extend to other binary CSPs. The difference between d-CNF-SAT and d-NAE-SAT shows that the type of constraints that one allows, affects the compressibility of the resulting CSP. The goal of this paper is to understand how the optimal compression size for a binary CSP depends on the type of legal constraints, with the aim of obtaining matching upper and lower bounds.

Before presenting our results, we give an example to illustrate our methods. Consider the NP-complete Exact d-CNF-Satisfiability (Exact d-SAT) problem, which asks whether there is a truth assignment that satisfies *exactly one* literal in each clause; the clauses have size at most d. While there are $\Theta(n^d)$ different clauses that can occur in an instance with n variables, the exact nature of the problem makes it possible to reduce any instance to an equivalent one with $n+1$ clauses. A clause such as $x_1 \vee x_3 \vee \neg x_5$ naturally corresponds to an equality constraint of the form $x_1 + x_3 + (1 - x_5) = 1$, since a 0/1-assignment to the variables satisfies exactly one literal of the clause if and only if it satisfies the equality. To find redundant clauses, transform each of the m clauses into an equality to obtain a system of equalities $A\mathbf{x} = \mathbf{b}$ where A is an $m \times n$ matrix, $\mathbf{x}$ is the column vector $(x_1, \ldots, x_n)$, and $\mathbf{b}$ is an integer column vector. Using Gaussian elimination, one can efficiently compute a basis B for the row space of the extended matrix $(A|b)$: a set of equalities such that every equality can be written as a linear combination of equalities in B. Since $(A|b)$ has $n+1$ columns, its rank is at most $n+1$ and the basis B contains at most $n+1$ equalities. To perform data reduction, remove all clauses from the Exact d-SAT instance whose corresponding equalities do not occur in B. If an assignment satisfies $f_1(\mathbf{x}) = b_1$ and $f_2(\mathbf{x}) = b_2$, then it also satisfies their sum $f_1(\mathbf{x}) + f_2(\mathbf{x}) = b_1 + b_2$, and any linear combination of the satisfied equalities. Since any equality not in B can be written as a linear combination of equalities in B, a truth assignment satisfying all clauses from B must necessarily also satisfy the remaining clauses, which shows the correctness of the data reduction procedure. The resulting instance can be encoded in $\mathcal{O}(n \log n)$ bits, as each of the remaining $n+1$ clauses has $d \in O(1)$ literals.

Our results

Our positive results are generalizations of the linear-algebraic data reduction tool for binary CSPs presented above. They reveal that the $\widetilde{\mathcal{O}}(n)$-bit compression for Exact d-SAT, the $\widetilde{\mathcal{O}}(n^{d-1})$-bit compression for d-NAE-SAT, and the $\mathcal{O}(n^d)$-bit compression for d-CNF-SAT

are samples of a gliding scale of problem complexity: more tightly constrained problems can be compressed better. We formalize this idea by considering a generic CSP whose constraints are of the form $f(\mathbf{x}) = 0$, where f is a bounded-degree polynomial and the constraint demands that $\mathbf{x}$ is a root of f. The example given earlier shows that Exact d-SAT can be expressed using degree-1 polynomials. We show that d-NAE-SAT and d-CNF-SAT can be expressed using equalities of polynomial expressions of degree $d-1$ and d. We study the following problem:

d-Polynomial root CSP	**Parameter:** The number of variables n.
Input: A list L of polynomial equalities over variables $V = \{x_1, \ldots, x_n\}$. An equality is of the form $f(x_1, \ldots, x_n) = 0$, where f is a multivariate polynomial of degree at most d.	
Question: Does there exist an assignment of the variables $\tau \colon V \to \{0, 1\}$ satisfying all equalities in L?	

Using a generalization of the argument presented above, the number of constraints in an instance of d-Polynomial root CSP can efficiently be reduced to $\mathcal{O}(n^d)$, even when the number of variables that occur in a constraint is not restricted. The latter implies, for example, that using degree-1 polynomials one can express the Exact sat problem with clauses of arbitrary size. When the number of variable occurrences in a constraint can be as large as n, it may take $\Omega(n)$ bits to encode a single constraint. After reducing the number of clauses in an Exact sat instance to $n+1$, one may therefore still require $\Theta(n^2)$ bits to encode the instance. This turns out to be unavoidable: we prove that Exact sat has no sparsification of size $\mathcal{O}(n^{2-\varepsilon})$ for any $\varepsilon > 0$, unless $\mathsf{NP} \subseteq \mathsf{coNP/poly}$. In general, we compress instances of d-Polynomial root CSP to bitsize $\widetilde{\mathcal{O}}(n^{d+1})$ when each constraint can be encoded in $\widetilde{\mathcal{O}}(n)$ bits. We prove that no compression to size $\mathcal{O}(n^{d+1-\varepsilon})$ is possible unless $\mathsf{NP} \subseteq \mathsf{coNP/poly}$. When each constraint can be encoded in $\widetilde{\mathcal{O}}(1)$ bits, the constraint reduction scheme reduces the size of an instance to $\widetilde{\mathcal{O}}(n^d)$. As we will show that d-NAE-SAT can be modeled using polynomials of degree $d-1$, this method strictly generalizes our earlier results [21] for d-NAE-SAT.

The linear-algebraic data reduction tool described above works over arbitrary fields F, allowing us to capture constraints such as "the number of satisfied literals in the clause is exactly two, when evaluated modulo 3". We therefore extend our study to the d-Polynomial root CSP problem over arbitrary fields F, and obtain similar positive and negative results.

Finally, we consider binary CSPs whose constraints are formed by *inequalities*, rather than equalities, of degree-d polynomials. This leads to the following generic problem:

d-Polynomial non-root CSP over F	**Parameter:** The number of variables n.
Input: A list L of polynomial inequalities over variables $V = \{x_1, \ldots, x_n\}$. An inequality is of the form $f(x_1, \ldots, x_n) \neq 0$, where f is a multivariate polynomial of degree $\leq d$.	
Question: Does there exist an assignment of the variables $\tau \colon V \to \{0, 1\}$ satisfying all inequalities in L?	

We present upper and lower bounds for problems of this type. When the polynomials are evaluated over a structure that is not a field, the situation changes significantly. For example, CSPs with constraints of the type "the number of satisfied literals in the clause is 1 or 2, when evaluated modulo 6" behave differently than the corresponding problem modulo 5, or modulo 7, because the integers modulo 6 do not form a field. Both our upper- and lower bound techniques fail when defining constraints with respect to composite moduli. We present connections to different areas of theoretical computer science where the distinction between prime and composite moduli plays a big role. More concretely, we show that obtaining polynomial sparsification upper bounds for d-Polynomial non-root CSP over the integers

modulo a composite, would resolve a long-standing problem concerning the representation of the OR-function using low-degree polynomials (cf. [2, 4, 29]).

Related work

Schaefer's Theorem [28] is a classic result relating the complexity of a binary CSP to the type of allowed constraints, separating the NP-complete from the polynomial-time solvable cases. A characterization of the kernelization complexity of min-ones CSPs parameterized by the number of variables was presented by Kratsch and Wahlström [25]. There are several parameterized complexity results for CSPs [8, 10, 24].

2 Preliminaries

A *parameterized problem* $\mathcal{Q}$ is a subset of $\Sigma^* \times \mathbb{N}$, where Σ is a finite alphabet. Let $\mathcal{Q}, \mathcal{Q}' \subseteq \Sigma^* \times \mathbb{N}$ be parameterized problems and let $h\colon \mathbb{N} \to \mathbb{N}$ be a computable function. A *generalized kernel for $\mathcal{Q}$ into $\mathcal{Q}'$ of size $h(k)$* is an algorithm that, on input $(x, k) \in \Sigma^* \times \mathbb{N}$, takes time polynomial in $|x| + k$ and outputs an instance (x', k') such that:

1. $|x'|$ and k' are bounded by $h(k)$, and
2. $(x', k') \in \mathcal{Q}'$ if and only if $(x, k) \in \mathcal{Q}$.

The algorithm is a *kernel* for $\mathcal{Q}$ if $\mathcal{Q} = \mathcal{Q}'$. It is a *polynomial (generalized) kernel* if $h(k)$ is a polynomial. Since a polynomial-time reduction to an equivalent sparse instance yields a generalized kernel, we use lower bounds for the sizes of generalized kernels to prove the non-existence of sparsification algorithms.

A *linear-parameter transformation* from a parameterized problem $\mathcal{Q}$ to a parameterized problem $\mathcal{Q}'$ is a polynomial-time algorithm that transforms any instance (x, k) of $\mathcal{Q}$ into an equivalent instance (x', k') of $\mathcal{Q}'$ such that $k' \in \mathcal{O}(k)$. It is easy to see (cf. [7]) that the existence of a linear-parameter transformation from $\mathcal{Q}$ to $\mathcal{Q}'$, together with a (generalized) kernel of size $\mathcal{O}(k^d)$ for $\mathcal{Q}'$, yields a generalized kernel of size $\mathcal{O}(k^d)$ for $\mathcal{Q}$. By contraposition, the existence of such a transformation implies that when $\mathcal{Q}$ does not have generalized kernels of size $\mathcal{O}(k^{d-\varepsilon})$, then $\mathcal{Q}'$ does not have generalized kernels of size $\mathcal{O}(k^{d-\varepsilon})$ either.

We use the framework of cross-composition [6] to establish kernelization lower bounds, requiring the definitions of polynomial equivalence relations and OR-cross-compositions. We repeat them here for completeness:

▶ **Definition 1** (Polynomial equivalence relation, [6, Def. 3.1])**.** An equivalence relation $\mathcal{R}$ on Σ^* is called a *polynomial equivalence relation* if the following conditions hold.

- There is an algorithm that, given two strings $x, y \in \Sigma^*$, decides whether x and y belong to the same equivalence class in time polynomial in $|x| + |y|$.
- For any finite set $S \subseteq \Sigma^*$ the equivalence relation $\mathcal{R}$ partitions the elements of S into a number of classes that is polynomially bounded in the size of the largest element of S.

▶ **Definition 2** (Cross-composition, [6, Def. 3.3])**.** Let $L \subseteq \Sigma^*$ be a language, let $\mathcal{R}$ be a polynomial equivalence relation on Σ^*, let $\mathcal{Q} \subseteq \Sigma^* \times \mathbb{N}$ be a parameterized problem, and let $f\colon \mathbb{N} \to \mathbb{N}$ be a function. An OR-*cross-composition of L into $\mathcal{Q}$* (with respect to $\mathcal{R}$) *of cost $f(t)$* is an algorithm that, given t instances $x_1, x_2, \ldots, x_t \in \Sigma^*$ of L belonging to the same equivalence class of $\mathcal{R}$, takes time polynomial in $\sum_{i=1}^{t} |x_i|$ and outputs an instance $(y, k) \in \Sigma^* \times \mathbb{N}$ such that:

- The parameter k is bounded by $\mathcal{O}(f(t) \cdot (\max_i |x_i|)^c)$, where c is some constant independent of t, and
- instance $(y, k) \in \mathcal{Q}$ if and only if there is an $i \in [t]$ such that $x_i \in L$.

▶ **Theorem 3** ([6, Theorem 6]). *Let $L \subseteq \Sigma^*$ be a language, let $\mathcal{Q} \subseteq \Sigma^* \times \mathbb{N}$ be a parameterized problem, and let d, ε be positive reals. If L is NP-hard under Karp reductions, has an* OR*-cross-composition into $\mathcal{Q}$ with cost $f(t) = t^{1/d+o(1)}$, where t denotes the number of instances, and $\mathcal{Q}$ has a polynomial (generalized) kernelization with size bound $\mathcal{O}(k^{d-\varepsilon})$, then* NP $\subseteq$ coNP/poly.

For $d \in \mathbb{N}$ we will refer to an OR-cross-composition of cost $f(t) = t^{1/d} \log(t)$ as a *degree-d cross-composition*. By Theorem 3, a degree-d cross-composition can be used to rule out generalized kernels of size $\mathcal{O}(k^{d-\varepsilon})$. Note that when studying sparsification, we use the number of vertices or variables in the instance (which is usually denoted by n) as the parameter value (which is usually denoted by k).

When interpreting truth assignments as elements of a field, we equate the value *true* with the 1 element in the field (multiplicative identity), and the value *false* with the 0 element (additive identity). Consequently, for a boolean variable x its negation $\neg x$ corresponds to $(1 - x)$. We let $\mathbb{Z}/m\mathbb{Z}$ denote the integers modulo m, which form a field if m is a prime number. The *degree* of a multivariate polynomial is the maximum degree of its monomials. Let $f(x_1, \ldots, x_d)$ be a d-variate polynomial over a field F. The *root set* of f is the algebraic variety $\{(e_1, \ldots, e_d) \in F^d \mid f(e_1, \ldots, e_d) = 0\}$. For a field F and a finite set $S \subseteq F$ of elements, the univariate polynomial $f(x) := \prod_{s \in S}(x - s)$ over F of degree $|S|$ has root set exactly S. We say that a field F is *efficient* if the field operations and Gaussian elimination can be done in polynomial time in the size of a reasonable input encoding. The field of rational numbers $\mathbb{Q}$, and all finite fields, are efficient. We use $[n]$ to denote $\{1, \ldots, n\}$. The $\widetilde{\mathcal{O}}$-notation suppresses polylogarithmic factors: $\widetilde{\mathcal{O}}(n) = \mathcal{O}(n \log^c n)$ for a constant c.

3 Kernel upper bounds

3.1 Polynomial root CSP

We start by showing how to reduce the number of constraints in instances of d-Polynomial root CSP, by extending the argument presented in the introduction.

▶ **Theorem 4.** *There is a polynomial-time algorithm that, given an instance (L, V) of d-*Polynomial root CSP *over an efficient field F, outputs an equivalent instance (L', V) with at most $n^d + 1$ constraints such that $L' \subseteq L$.*

Proof. Given a list L of polynomial equalities over variables V for d-Polynomial root CSP, we use linear algebra to find redundant constraints. Observe that $(x_i)^c = x_i$ for all 0/1-assignments and $c \in \mathbb{N}_+$. As constraints are evaluated over 0/1-assignments, we may assume without loss of generality that the monomials in each of the polynomials are multilinear: each monomial consists of a coefficient from F multiplied by distinct variables.

Create a matrix A with $|L|$ rows and a column for every multilinear monomial of degree at most d over variables from V. Let position $a_{i,j}$ in A be the coefficient of the monomial corresponding to column j in the polynomial equality corresponding to row i.

Compute a basis B of the row space of matrix A, for example using Gaussian elimination [18], and let L' consist of the equalities in L whose corresponding row appears in the basis. Since $L' \subseteq L$, it follows that if the original instance has a satisfying assignment, the reduced instance has a satisfying assignment as well. The crucial part of the correctness proof is to establish the converse.

▶ **Claim 5.** *If an assignment $\tau \colon V \to \{0, 1\}$ of the variables in V satisfies the equalities in L', then it satisfies all equalities in L.*

Proof. Consider any equality $(f(\mathbf{x}) = 0) \in L \setminus L'$, since equalities in L' are trivially satisfied, and assume it corresponds to the i'th matrix row. Let $f_j(\mathbf{x})$ be the polynomial represented in the j'th row of matrix A for $j \in [|L|]$. Without loss of generality, let the basis of A correspond to its first m rows $\mathbf{a}_1, \ldots, \mathbf{a}_m$. We then have $i > m$, and by the definition of basis there exist $\beta_1, \ldots, \beta_m \in F$ such that $\mathbf{a}_i = \sum_{j=1}^{m} \beta_j \mathbf{a}_j$. Let $\mathbf{t}$ be the column vector containing, for each multilinear monomial of degree $\leq d$ in variables $x_1, \ldots, x_n$, the evaluation under τ. For example, for monomial $x_1 x_3$ it contains $\tau(x_1) \cdot \tau(x_3)$. By using the same order of monomials as in the construction of A, we obtain for all $j \in [|L|]$ that $f_j(\tau(x_1), \ldots, \tau(x_n)) = \mathbf{a}_j \mathbf{t}$, the inner product of $\mathbf{a}_j$ and $\mathbf{t}$. It follows that $\mathbf{a}_j \mathbf{t} = 0$ for all $j \in [m]$, since satisfying L' implies $f_j(\tau(x_1), \ldots, \tau(x_n)) = 0$. Now observe that

$$f_i(\mathbf{x}) = \mathbf{a}_i \mathbf{t} = \sum_{j=1}^{m} (\beta_j \mathbf{a}_j) \mathbf{t} = \sum_{j=1}^{m} \beta_j (\mathbf{a}_j \mathbf{t}) = \sum_{j=1}^{m} \beta_j \cdot 0 = 0,$$

which proves the claim. ⌟

▶ **Claim 6.** *The number of constraints in the resulting kernel is bounded by $n^d + 1$.*

Proof. The size of a basis of any matrix over a field equals its rank, which is bounded by the number of columns. As there is a column for each multilinear monomial of degree at most d, there are at most $\sum_{i=0}^{d} \binom{n}{i}$ constraints in the basis. Now observe that $\sum_{i=1}^{d} \binom{n}{i} \leq n^d$. The left side counts nonempty subsets of $[n]$ of size at most d, each of which can be mapped to a distinct d-tuple by repeating an element. Since there are n^d d-tuples, the claim follows. ⌟

This concludes the proof of Theorem 4. ◀

When each constraint can be encoded in $\widetilde{\mathcal{O}}(n)$ bits, for example when each polynomial can be represented as an arithmetic circuit of size $\mathcal{O}(n)$, Theorem 4 gives a kernelization of size $\widetilde{\mathcal{O}}(n^{d+1})$. When constraints can be encoded in $\widetilde{\mathcal{O}}(1)$ bits, which may occur when constraints have constant arity, we obtain kernels of bitsize $\widetilde{\mathcal{O}}(n^d)$. For explicit examples consider the following problem, where optionally a prime p may be chosen.

Generalized d-Sat (mod p)	**Parameter:** The number of variables n
Input: A set of clauses $\mathcal{C}$ over variables $V := \{x_1, \ldots, x_n\}$, and for each clause a set $S_i \subset \mathbb{N} \cup \{0\}$ with $\lvert S_i \rvert \leq d$. Each clause is a set of distinct literals of the form x_i or $\neg x_i$. **Question:** Does there exist a truth assignment for the variables V such that the number of satisfied literals in clause i lies in $S_i \pmod p$ for all i?	

▶ **Corollary 7.** Generalized d-Sat *and* Generalized d-Sat mod p *both have a kernel with $n^d + 1$ clauses that can be encoded in $\mathcal{O}(n^{d+1} \log n)$ bits.*

Proof. To reduce the number of clauses using Theorem 4, we only have to provide a polynomial of degree at most d to represent each constraint. Consider a clause involving k variables $x_{i_1}, \ldots, x_{i_k}$. Let $t_j = x_{i_j}$ if variable x_{i_j} occurs positively in the clause, and let $t_j = (1 - x_{i_j})$ if the variable occurs negatively. Then the number of satisfied literals in the clause is given by the degree-1 polynomial $f(x_{i_1}, \ldots, x_{i_k}) := \sum_{i=1}^{k} t_i$. Let $F(x)$ be a polynomial with root set $S_j \pmod p$ of degree at most $|S_j|$. We obtain $F(f(\mathbf{x})) \equiv 0 \pmod p$ if and only if $\mathbf{x}$ satisfies the clause. Note that the degree of $F(f(\mathbf{x}))$ is at most $|S_j| \leq d$.

Applying Theorem 4 to the resulting instance of d-Polynomial root CSP identifies a subset of at most $n^d + 1$ constraints which preserve the answer to the Sat problem. Each clause contains at most $2n$ literals, which can be encoded in $\mathcal{O}(\log n)$ bits each. Additionally,

for each clause we need to store the set S_i of at most d integers, which have value at most $2n$ in relevant inputs. As d is a constant, the instance can be encoded in $\mathcal{O}(n^{d+1}\log n)$ bits. ◀

Corollary 7 yields a new way to get a nontrivial compression for d-NAE-SAT, which is conceptually simpler than the existing approach which requires an unintuitive lemma by Lovász [27]. The new approach gives the same size bound as given earlier [21].

▶ **Corollary 8.** *d-NAE-SAT has a kernel with $n^{d-1}+1$ clauses and bitsize $\mathcal{O}(n^{d-1}\log n)$.*

Proof. A clause of size $k \leq d$ is not-all-equal satisfied if and only if the number of satisfied literals lies in $S := \{1,\ldots,k-1\}$. Using Corollary 7 we can reduce the number of clauses to $n^{d-1}+1$. Each clause has $d \in O(1)$ variables and can thus be encoded in $\mathcal{O}(\log n)$ bits. ◀

3.2 Polynomial non-root CSP

In this section we consider d-Polynomial non-root CSP. In Section 4.2 we will show that, over the field of rational numbers, the problem cannot be compressed to size polynomial in n, unless NP $\subseteq$ coNP/poly. We therefore consider the field $\mathbb{Z}/p\mathbb{Z}$ of integers modulo a prime p.

▶ **Theorem 9.** *There is a polynomial-time algorithm that, given an instance (L,V) of d-Polynomial non-root CSP over $\mathbb{Z}/p\mathbb{Z}$, outputs an equivalent instance (L',V) with $\mathcal{O}(n^{d(p-1)})$ constraints such that $L' \subseteq L$.*

Proof. Suppose we are given a list of polynomial inequalities L over variables V. Observe that an inequality $f(\mathbf{x}) \not\equiv 0 \pmod p$ is equivalent to $f(\mathbf{x}) \in \{1,\ldots,p-1\} \pmod p$.

Let $F\colon \mathbb{Z}/p\mathbb{Z} \to \mathbb{Z}/p\mathbb{Z}$ be a polynomial of degree $p-1$ with root set $\{1,\ldots,p-1\}$ modulo p, which exists since $\mathbb{Z}/p\mathbb{Z}$ is a field. Then $f(\mathbf{x}) \not\equiv 0 \pmod p$ can equivalently be stated as $F(f(\mathbf{x})) \equiv 0 \pmod p$. It is easy to see that $F(f(\mathbf{x}))$ is a polynomial of degree at most $d(p-1)$. Therefore, L can be written as an instance of $d(p-1)$-Polynomial root CSP by replacing every polynomial f by $F \circ f$. By Theorem 4, the proof follows. ◀

In Section 4.2 we will establish a nearly-matching lower bound counterpart to Theorem 9.

4 Kernel lower bounds

4.1 Polynomial root CSP

We now turn our attention to lower bounds, starting with d-Polynomial root CSP over $\mathbb{Q}$. We start by proving that Exact Red-Blue Dominating Set does not have generalized kernels of bitsize $\mathcal{O}(n^{2-\varepsilon})$ for any $\varepsilon > 0$, unless NP $\subseteq$ coNP/poly. The same lower bound for 1-Polynomial root CSP will follow by a linear-parameter transformation. We then show how to generalize this result to d-Polynomial root CSP. As a starting problem for the cross-composition we will use the NP-hard Red-Blue Dominating Set (RBDS) [12, 22].

Red-Blue Dominating Set (RBDS)	**Parameter:** The number of vertices n
Input: A bipartite graph $G = (R \cup B, E)$ containing red (R) and blue (B) vertices, and an integer k.	
Question: Does there exist a set $D \subseteq R$ with $\lvert D\rvert \leq k$ such that every vertex in B has at least one neighbor in D?	

Exact Red Blue Dominating Set (ERBDS) is defined similarly, except that every vertex in B must have *exactly one* neighbor in D. Furthermore we will not bound the size of such a set, but merely ask for the existence of any ERBDS.

▶ **Theorem 10.** Exact Red-Blue Dominating Set *parameterized by the number of vertices n does not have a generalized kernel of size $\mathcal{O}(n^{2-\varepsilon})$, unless* NP $\subseteq$ coNP/poly.

Proof. We will prove this result by giving a degree-2 cross-composition from RBDS to ERBDS. We start by giving a polynomial equivalence relation $\mathcal{R}$ on inputs of RBDS. Let two instances of RBDS be equivalent under $\mathcal{R}$ if they have the same number of red vertices, the same number of blue vertices, and the same maximum size of a RBDS. It is easy to check that $\mathcal{R}$ is a polynomial equivalence relation.

Assume we are given t instances of RBDS, labeled $X_{i,j}$ for $i, j \in [\sqrt{t}]$, from the same equivalence class of $\mathcal{R}$. If the number of instances given is not a square, we duplicate one of the input instances until a square number is reached. Since this changes the number of inputs by at most a factor four, this does not influence the cross-composition. Instance $X_{i,j}$ consists of graph $G_{i,j}$ with a set of red vertices $R_{i,j}$ and blue vertices $B_{i,j}$. Call the number of red vertices in every instance m_R, the number of blue vertices m_B, and the required size of the dominating set k. For each instance enumerate the red vertices as $r_1, \ldots, r_{m_R}$ and the blue vertices as $b_1, \ldots, b_{m_B}$, arbitrarily. Create instance G' for ERBDS by the following steps. Figure 1 shows a sketch of G'.

1. Create $\sqrt{t}$ sets $U_1, \ldots, U_{\sqrt{t}}$ each consisting of $k \cdot m_R$ red vertices, such that for all $\ell \in [\sqrt{t}]$ $U_\ell := \{u^\ell_{i,j} \mid i \in [k],\ j \in [m_R]\}$.
2. Similarly create $\sqrt{t}$ sets $V_1, \ldots, V_{\sqrt{t}}$, each consisting of $k \cdot m_B$ blue vertices, and define $V_\ell := \{v^\ell_{i,j} \mid i \in [k], j \in [m_B]\}$ for all $\ell \in [\sqrt{t}]$.
3. For each $i \in [k]$ add the edge from $u^\ell_{i,j}$ to $v^{\ell'}_{i,j'}$ if $\{r_j, b_{j'}\}$ is an edge in instance $X_{\ell,\ell'}$ with $\ell, \ell' \in [\sqrt{t}], j \in [m_R], j' \in [m_B]$.

By steps 1 to 3, the graph induced by the vertices in $U_\ell \cup V_{\ell'}$ consists of k vertex-disjoint copies of $G_{\ell,\ell'}$. The next steps are used to ensure that there are exactly k vertices from U in any ERBDS, which must all belong to the same set U_ℓ.

4. Create k blue vertices $W := \{w_i \mid i \in [k]\}$ and connect all vertices $\{u^\ell_{i,j} \mid j \in [m_R], \ell \in [\sqrt{t}]\}$ to w_i for $i \in [k]$.
5. Create blue vertices d^ℓ_i for $\ell \in [\sqrt{t}]$ and $i \in [k]$. Connect vertex d^ℓ_i to the vertices $u^\ell_{i,j}$ with $j \in [m_R]$. Add blue vertex S and red vertices $Z := \{z_j \mid j \in [\sqrt{t}]\}$ and connect z_j to d^ℓ_i for $i \in [k]$ and $\ell \neq j \in [\sqrt{t}]$. Connect all vertices in Z to vertex S.

The next steps ensure that some of the blue vertices in one set V_ℓ need to be dominated by vertices from U, while all other vertices in V can be dominated "for free".

6. Add sets of gadgets C_ℓ for $\ell \in [\sqrt{t}]$. Each set consists of $m_B \cdot k$ selector gadgets $c^\ell_{i,j}$. A selector gadget consists of $k+1$ red vertices labeled $a_1, \ldots, a_{k+1}$ that are all connected to a blue vertex b that is private to the gadget. Furthermore, in gadget $c^\ell_{i,j}$, the vertex a_x for $x \in [k]$ is connected to $v^\ell_{x,j}$ for $j \in [m_B]$, $\ell \in [\sqrt{t}]$ and $i \in [k]$. By this construction an ERBDS uses at most one red vertex from each gadget, to dominate one vertex from V.
7. Add red vertices $Y := y_1, \ldots, y_{\sqrt{t}}$ and connect y_ℓ to the blue vertices of gadgets $c^\ell_{1,j}$ for all $j \in [m_B]$, $\ell \in [\sqrt{t}]$. Connect $y_1, \ldots, y_{\sqrt{t}}$ to the new blue vertex S'.

This concludes the construction of graph G', with red vertices ($U \cup Y \cup Z \cup$ vertices labeled $a_1, \ldots, a_{k+1}$ in C), and blue vertices ($V \cup D \cup \{S, S'\} \cup$ vertices labeled b in C).

▶ **Claim 11.** *For any* ERBDS E *of G', there exists an index $\ell \in [\sqrt{t}]$ such that $U_x \cap E = \emptyset$ for all $x \neq \ell \in [\sqrt{t}]$ and $|E \cap \{u^\ell_{i,j} \mid j \in [m_R]\}| = 1$ for all $i \in k$.*

Proof. By Step 5, blue vertex S has neighborhood $\{z_\ell \mid \ell \in [\sqrt{t}]\}$. Exactly one of these vertices is contained in E; let this be z_ℓ. The neighborhood of z_ℓ contains $\{d^j_i \mid i \in [k], j \in [\sqrt{t}] \setminus \{\ell\}\}$. Thereby, no other neighbors from vertices in this set can be in E, implying no vertices from U_i for $i \neq \ell \in [\sqrt{t}]$ can be in E. In other words, $U_i \cap E = \emptyset$ for all $i \neq \ell \in [\sqrt{t}]$.

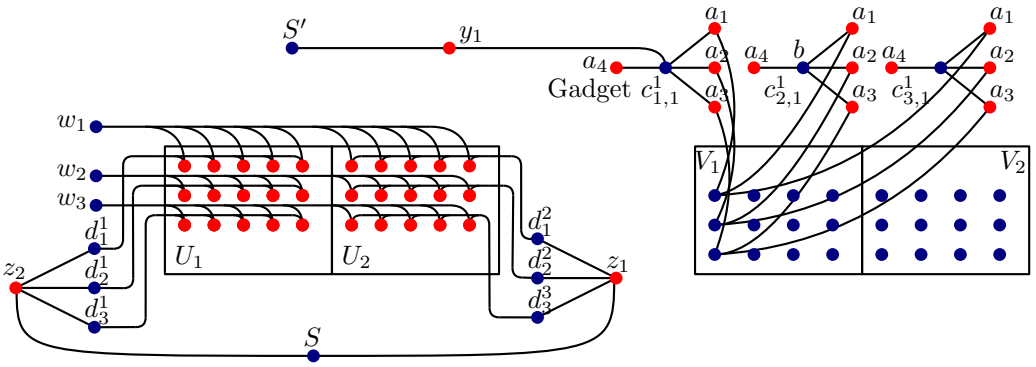

Figure 1 The graph G' created in the proof of Theorem 10, for $k = 3$, $m_R = 5$, $m_B = 4$, and $t = 4$. Edges between U and V are left out for simplicity. Of the 24 gadgets in C only $c^1_{1,1,}$, $c^1_{2,1}$, and $c^1_{3,1}$ are shown. Vertex y_2 is left out.

By Step 4, the neighborhood of blue vertex w_i for $i \in [k]$ is exactly $\{u^x_{i,j} \mid x \in [\sqrt{t}], j \in [m_R]\}$. It follows that exactly one vertex in this set is in E for all i. By the previous argument the vertex cannot be from U_x for $x \neq \ell$, hence it is from U_ℓ. ⌟

▶ **Claim 12.** *For any* ERBDS E *of* G'*, there exists* ℓ *such that* $E \cap c^\ell_{1,j} = \emptyset$ *for all* $j \in [m_B]$.

Proof. By Step 7, blue vertex S' has neighborhood $\{y_\ell \mid \ell \in [\sqrt{t}]\}$. Exactly one of these vertices is contained in E; let this be y_ℓ. It is connected to the blue vertex of all gadgets $c^\ell_{1,j}$ for $j \in [m_B]$. Since all red vertices in a gadget $c^\ell_{1,j}$ for $j \in [m_B]$ have a blue neighbor b that is also adjacent to $y_\ell \in E$, the red vertices in these gadgets are not present in E. ⌟

▶ **Claim 13.** *For any* ERBDS E *of* G'*, there exists an index* ℓ *such that for every* $j \in [m_B]$ *at least one of the vertices in* $\{v^\ell_{i,j} \mid i \in [k]\}$ *has a neighbor in* $E \cap U$.

Proof. By Claim 12 there exists $\ell \in [\sqrt{t}]$ such that $E \cap c^\ell_{1,j} = \emptyset$ for all $j \in [m_B]$. Consider an arbitrary $j \in [m_B]$. The k vertices in $\{v^\ell_{i,j} \mid i \in [k]\}$ are connected to k gadgets $c^\ell_{1,j}, c^\ell_{2,j}, \ldots, c^\ell_{k,j}$, and to some vertices in U. From each gadget, at most one red vertex is in E, since the red vertices have a common blue neighbor. Any red gadget vertex is connected to only one vertex in V. Since no vertex of gadget $c^\ell_{1,j}$ is in E, at most $k-1$ of the vertices in $\{v^\ell_{i,j} \mid i \in [k]\}$ have a neighbor in $E \cap C_\ell$. Consequently, at least one of these vertices has a neighbor in $E \cap U$ for each $j \in [m_B]$. ⌟

▶ **Claim 14.** *If* G' *has an* ERBDS*, then some input* $X_{i,j}$ *has a* RBDS *of size at most* k.

Proof. Assume G' has an ERBDS, say E. By Claim 13, there exists $\ell_2 \in [\sqrt{t}]$, such that for every $j \in [m_B]$ at least one of the vertices in $\{v^{\ell_2}_{i,j} \mid i \in [k]\}$ has a neighbor in $E \cap U$. By Claim 11, there exists $\ell_1 \in [\sqrt{t}]$ with $U_i \cap E = \emptyset$ for all $i \neq \ell_1$, so these neighbors lie in U_{ℓ_1}.

We now construct a RBDS E' for instance X_{ℓ_1,ℓ_2}. For each $j \in [m_R]$, add r_j to E' if $E \cap \{u^{\ell_1}_{i,j} \mid i \in [k]\} \neq \emptyset$. By Claim 11, it follows that E' has size at most k, as required. It remains to show that every vertex in B_{ℓ_1,ℓ_2} has a neighbor in E'. If some vertex b_j from B_{ℓ_1,ℓ_2} does not have a neighbor in E', then none of the vertices $\{v^{\ell_2}_{i,j} \mid i \in [k]\}$ have a neighbor in $E \cap U_{\ell_1}$. This contradicts Claim 13. Hence E' is an RBDS of size at most k for B_{ℓ_1,ℓ_2}. ⌟

▶ **Claim 15.** *If some input instance has a* RBDS *of size at most* k*, then* G' *has an* ERBDS.

Proof. Suppose instance X_{ℓ_1,ℓ_2} has a RBDS E' of size k consisting of vertices $r_{i_1}, \ldots, r_{i_k} \subseteq R_{\ell_1,\ell_2}$. We construct an ERBDS E for G'. Start by choosing vertices $u^{\ell_1}_{x,i_x}$ for $x \in [k]$, so for every vertex in E' we pick one vertex in the ERBDS for G'. Furthermore we choose the red vertices z_{ℓ_1} and y_{ℓ_2} to be in E. To exactly dominate the blue vertices in V, we use the gadgets in C as follows. For $\ell \neq \ell_2 \in [\sqrt{t}]$, add red vertex a_x of gadget $c^{\ell}_{x,j}$ if vertex $v^{\ell}_{x,j}$ does not yet have a neighbor in E, for $j \in [m_R]$. Else, add vertex a_{k+1} of gadget $c^{\ell}_{x,j}$ to E, in order to exactly dominate the blue vertex of this gadget.

To exactly dominate the vertices in V_{ℓ_2} we apply a similar procedure, except that gadget $c^{\ell}_{1,j}$ cannot be used since its blue vertex b is already dominated by y_{ℓ_2}. Since E' is a RBDS of instance X_{ℓ_1,ℓ_2}, for each $j \in [m_B]$ at least one vertex from set $\{v^{\ell_2}_{i,j} \mid i \in [k]\}$ has a neighbor in $E \cap U$. As such, the $k-1$ remaining gadgets can be used to each dominate one of the $k-1$ remaining vertices in this set, if they do not already have a neighbor in $E \cap U$. If no red vertex of a gadget is needed to dominate, we choose vertex a_{k+1} of the gadget in E to dominate the blue vertex in the gadget.

It is straight-forward to verify that this results in an ERBDS for G'. ⌟

From Claims 14 and 15 it follows that graph G' has an ERBDS if and only if at least one of the input instances has a RBDS of size at most k. The graph G' has $\mathcal{O}(\sqrt{t} \cdot (m_R + m_B)^3)$ vertices, which is suitably bounded for a cross-composition. By Theorem 3, it follows that ERBDS parameterized by the number of vertices n does not have a generalized kernel of size $\mathcal{O}(n^{2-\varepsilon})$ for any $\varepsilon > 0$, unless $\mathsf{NP} \subseteq \mathsf{coNP/poly}$. ◀

Using Theorem 10 we provide lower bounds for constraint satisfaction problems.

▶ **Corollary 16.** *The problems* EXACT SATISFIABILITY *and* 1-POLYNOMIAL ROOT CSP *over* $\mathbb{Q}$*, parameterized by the number of variables* n*, do not have a generalized kernel of size* $\mathcal{O}(n^{2-\varepsilon})$ *for any* $\varepsilon > 0$*, unless* $\mathsf{NP} \subseteq \mathsf{coNP/poly}$.

Proof. By Theorem 10 and the discussion in Section 2, it suffices to give linear-parameter transformations from ERBDS parameterized by the number of vertices to the two mentioned problems. Consider an instance $G = (R \cup B, E)$ of ERBDS. Create a binary variable x_r for each $r \in R$. For each blue vertex $b \in B$ create a clause of the form $\bigvee_{r \in N(b)} x_r$ (to build an instance of EXACT SAT), or create a constraint $\sum_{r \in N(b)} x_r = 1$ (to build an instance of CSP). The resulting instance has a satisfying 0/1-assignment if and only if G has an ERBDS. Since the number of variables is $|R| \leq n$, these are valid linear-parameter transformations. ◀

▶ **Theorem 17.** d-POLYNOMIAL ROOT CSP *over* $\mathbb{Q}$ *parameterized by the number of variables* n *does not have a generalized kernel of size* $\mathcal{O}(n^{d+1-\varepsilon})$ *for any* $\varepsilon > 0$*, unless* $\mathsf{NP} \subseteq \mathsf{coNP/poly}$.

Proof. The case $d = 1$ is covered by Corollary 16; we consider $d \geq 2$ and give a degree-$(d+1)$ cross-composition from RBDS, re-using some parts of the proof of Theorem 10. Suppose we are given $t = r^{d+1}$ instances of RBDS from the same equivalence class of $\mathcal{R}$, all having m_R red vertices and m_B blue vertices. By a similar padding argument as before, we may assume r is an integer. Split the inputs into groups of size r^2 and apply the cross-composition of Theorem 10 to each group, followed by the linear-parameter transformation in Corollary 16. We obtain r^{d-1} instances of 1-POLYNOMIAL ROOT CSP with $O(r \cdot \text{poly}(m_R + m_B))$ variables each, such that the answer to each composed instance is the logical OR of the answers to the RBDS instances in its group. Label the instances resulting from the group compositions $X_{i_1,\ldots,i_{d-1}}$ with $i_1, \ldots, i_{d-1} \in [r]$. They all use the same number of variables; label the variables in each instance as $x_1, \ldots, x_q$. Create an instance L' of d-POLYNOMIAL ROOT CSP as follows:

1. Create variables $x_1, \ldots, x_q$. Create sets $Y_1, \ldots, Y_{d-1}$ of r variables each, where $Y_i := \{y_j^i \mid j \in [r]\}$. Add the requirement $\sum_{j \in [r]} y_j^i = 1$ to L' for each $i \in [d-1]$.
2. Let the list of equations of instance $X_{i_1,\ldots,i_{d-1}}$ be $L_{i_1,\ldots,i_{d-1}}$. For every equality $f(\mathbf{x}) = 1$ in $L_{i_1,\ldots,i_{d-1}}$ with $i_1, \ldots, i_d \in [r]$, add the following equality to L':

$$f(\mathbf{x}) \cdot \prod_{z \in [d-1]} y_{i_z}^z = \prod_{z \in [d-1]} y_{i_z}^z.$$

The polynomial equalities have degree $\leq d$ since $f(\mathbf{x})$ has degree 1. The number of variables is $q + (d-1) \cdot r \in \mathcal{O}(r \cdot d \cdot \text{poly}(m_R + m_B)) \in \mathcal{O}(t^{1/(d+1)}\text{poly}(m_R + m_B))$. It remains to show that L' is satisfiable if and only if one of the input instances has an ERBDS. Since Theorem 10 gives a correct cross-composition, it is sufficient to show that L' is satisfiable if and only if one of the r^{d-1} instances of 1-Polynomial root CSP has a solution.

($\Rightarrow$) Suppose L' is satisfied by some assignment. Then from each Y_i for $i \in [d-1]$, exactly one variable is set to 1. So suppose variables $y_{i_z}^z$ are set to 1 for $z \in [d-1], i_z \in [r]$. Then from instance $X_{i_1,\ldots,i_{d-1}}$, all polynomial equations are copied to L' and multiplied by 1 on both sides. Hence they are satisfied by the assignment to $\mathbf{x}$.

($\Leftarrow$) Suppose instance $X_{i_1,\ldots,i_{d-1}}$ of 1-Polynomial root CSP has a satisfying assignment. Set the $\mathbf{x}$-variables according to this assignment. Furthermore, set variables $y_{i_z}^z$ for $z \in [d-1]$ to 1, set all other variables to 0. Thereby the sum of variables in each set Y_i is 1, as required. Furthermore, any equation added in Step 2 of the construction is satisfied in the following way. If it belongs to instance $X_{i_1,\ldots,i_{d-1}}$, it is satisfied by definition. Equations belonging to any other instance are trivially satisfied since both sides are multiplied by zero. ◀

Observe that the polynomials constructed in Theorem 17 have a simple form: each polynomial is a product of $(d-1)$ Y-variables multiplied by a sum of distinct variables from $\mathbf{x}$. Each polynomial can therefore be encoded in $\widetilde{\mathcal{O}}(n)$ bits, where n is the number of variables in the constructed CSP. The sparsification of Theorem 4 therefore encodes such instances in $\widetilde{\mathcal{O}}(n^{d+1})$ bits. The lower bound shows that this is optimal up to $n^{o(1)}$ factors.

We expect the lower bound of Theorem 17 to extend to arbitrary finite fields of prime order, except for the case $d = 1$ over $\mathbb{Z}/2\mathbb{Z}$, which is polynomial-time solvable [28].

4.2 Polynomial non-root CSP

We start our lower bound discussion for d-Polynomial non-root CSP by considering polynomials over $\mathbb{Q}$. 1-Polynomial non-root CSP over $\mathbb{Q}$ does not have a generalized kernel of size bounded by any polynomial in n, unless $\mathsf{NP} \subseteq \mathsf{coNP/poly}$. This follows from the fact that CNF-Satisfiability parameterized by the number of variables does not have a kernel of size polynomial in n unless $\mathsf{NP} \subseteq \mathsf{coNP/poly}$ [11, 15], together with the fact that a clause such as $(x_1 \vee \neg x_3 \vee x_4)$ is satisfied by a 0/1-assignment if and only if $x_1 + (1 - x_3) + x_4 \neq 0$ over $\mathbb{Q}$. In the remainder of the section we investigate the behavior over finite fields.

In Theorem 9 we provided a kernel for d-Polynomial non-root CSP over $\mathbb{Z}/p\mathbb{Z}$ for primes p. It is natural to ask whether similar results can be obtained when working with polynomials modulo an arbitrary integer m. When m is composite, our kernelization fails. We can show that this is not a shortcoming of our proof strategy, but a necessity due to the fact that constraints expressed by equalities of degree-d polynomials modulo composite numbers can model more complex constraints than degree-d polynomials modulo a prime. For example, it is known (cf. [1, §2]) that there is a degree-3 polynomial f over the integers modulo 6 which represents a logical OR of size 27 in the following way:

$$f(x_1, \ldots, x_{27}) \not\equiv 0 \pmod 6 \Leftrightarrow (x_1 \vee \ldots \vee x_{27}). \tag{1}$$

By this expressibility of a size-27 OR by a polynomial of degree 3 over $\mathbb{Z}/6\mathbb{Z}$ using the same variables, it is easy to give a linear-parameter transformation from 27-CNF-SAT to 3-POLYNOMIAL NON-ROOT CSP (mod 6). Using known lower bounds for d-CNF-SAT [11, Theorem 1], this implies the latter problem has no kernel of $\mathcal{O}(n^{27-\varepsilon})$ bits, unless $\mathsf{NP} \subseteq \mathsf{coNP/poly}$. Plugging in the degree of 3 and modulus 6 into the bound of Theorem 9 would give a reduction to $\mathcal{O}(n^{3\cdot(6-1)}) = \mathcal{O}(n^{15})$ constraints and would contradict the lower bound. The example therefore shows that the problem is more complex for composite moduli.

For more general non-primes, we can prove a lower bound using a general construction by Bhowmick *et al.* [4] of low-degree polynomials representing OR in the sense of Equation 1.

▶ **Theorem 18.** *Let m be a non-prime with a prime factorization consisting of r distinct primes, such that $m = \prod_{i\in[r]} p_i$. Then d-*POLYNOMIAL NON-ROOT CSP *(mod m) parameterized by the number of variables n does not have a generalized kernel of size $\mathcal{O}(n^{(d^r)/2-\varepsilon})$ for any $\varepsilon > 0$, unless* $\mathsf{NP} \subseteq \mathsf{coNP/poly}$.

Proof. Bhowmick *et al.* [4, Appendix A] provide a way to find a polynomial f of degree $2\lceil N^{1/r}\rceil$ such that for all $x_1, \ldots, x_N \in \{0,1\}$,

$$f(x_1, \ldots, x_N) \not\equiv 0 \pmod{m} \Leftrightarrow (x_1 \vee \cdots \vee x_N). \tag{2}$$

This implies that for $N = d^r/2$, we can find a polynomial f of degree d satisfying the above equation. As such, d-POLYNOMIAL NON-ROOT CSP can express a logical OR of size $d^r/2$ using the same variables. In this way we can give a linear-parameter transformation from $(d^r/2)$-CNF-SAT to d-POLYNOMIAL NON-ROOT CSP. By [11, Theorem 1], the latter problem does not have a generalized kernel of size $\mathcal{O}(n^{(d^r)/2-\varepsilon})$ for any $\varepsilon > 0$, unless $\mathsf{NP} \subseteq \mathsf{coNP/poly}$. ◀

In case m does not have a prime factorization in which all primes are distinct, it is possible to obtain weaker a lower bound using a result by Barrington *et al.* [2], which proves that there exists a polynomial of degree $\mathcal{O}(\ell N^{1/r})$ that represents a logical OR when taken modulo m. Here ℓ is the largest prime factor of m. For prime moduli, we provide a lower bound almost matching the upper bound in Section 3.2.

▶ **Theorem 19.** *Let p be a prime. d-*POLYNOMIAL NON-ROOT CSP *(mod p) parameterized by the number of variables n does not have a generalized kernel of size $\mathcal{O}(n^{d(p-1)-\varepsilon})$ for any $\varepsilon > 0$, unless* $\mathsf{NP} \subseteq \mathsf{coNP/poly}$.

Proof. We aim at giving a linear-parameter transformation from $d(p-1)$-CNF-SAT. We proceed similarly as in the proof of Theorem 18. It is known (cf. [3, Theorem 24]) that for each prime p and integer d, there is a polynomial f of degree d modulo p, such that for any $x_1, \ldots, x_{d(p-1)} \in \{0,1\}$ we have:

$$f(x_1, \ldots, x_{d(p-1)}) \not\equiv 0 \pmod{p} \Leftrightarrow (x_1 \vee x_2 \vee \cdots \vee x_{d(p-1)}).$$

This allows the linear-parameter transformation to be carried out as in Theorem 18. For completeness, we repeat one way of constructing such polynomials here.

For a set of 0/1 variables S let $\text{OR}(S)$ be 0 if all variables in S are set to *false* and 1 otherwise. Let X be the set of $d(p-1)$ variables in a certain clause. Divide X into $p-1$ groups of size d, labeled $X_1, \ldots, X_{p-1}$. For each group X_i for $i \in [p-1]$ we will construct one d-variate polynomial g of degree d, such that $g(X_i) = \text{OR}(X_i)$, implying this polynomial always outputs either 0 or 1, given 0/1-valued arguments. Define for $i \in [d]$ the polynomial g_i to represent the logical OR on i variables, such that

$$g_1(x_1) := x_1,$$

clearly $g_1(x_1) = \text{OR}(x_1)$ for all $x_1 \in \{0,1\}$. Recursively, we define for $i \geq 2$

$$g_i(x_1, \ldots, x_{i-1}, x_i) := x_i + (1 - x_i) \cdot g_{i-1}(x_1, \ldots, x_{i-1})$$

and we let $g := g_d$. It is easy to see that each g_i corresponds to a logical OR of i variables and has degree at most i. As such, g has degree at most d.

Now let

$$f(X) := \sum_{i \in [p-1]} g(X_i),$$

such that $f(X) \equiv 0 \pmod{p}$ if and only if all variables in X are *false*, for $X \in \{0,1\}^{d(p-1)}$. Since every polynomial g has degree at most d, it follows that f has degree at most d. ◀

5 Conclusion

We have given upper and lower bounds on the kernelization complexity of binary CSPs that can be represented by polynomial (in)equalities, obtaining tight sparsification bounds in several cases. Our main conceptual contribution is to analyze constraints on binary variables based on the minimum degree of multivariate polynomials whose roots, or non-roots, capture the satisfying assignments. The ultimate goal of this line of research is to characterize the optimal sparsification size of a binary CSP based on easily accessible properties of the constraint language. To reach this goal, several significant hurdles have to be overcome.

For d-Polynomial non-root CSP (mod 6), we do not know of any way to reduce the number of constraints to polynomial in n. This difficulty is connected to longstanding questions regarding the minimum degree of a multivariate polynomial modulo 6 that represents the OR-function of n variables in the sense of Equation 1. As exploited in the construction of Theorem 18, if the OR-function with $g(d)$ inputs can be represented by polynomials of degree d, then d-Polynomial non-root CSP cannot be compressed to size $\mathcal{O}(n^{g(d)-\varepsilon})$ unless NP $\subseteq$ coNP/poly. By contraposition, a kernelization with size bound $\widetilde{\mathcal{O}}(n^{h(d)})$ implies a lower bound of $h^{-1}(d)$ on the degree of a polynomial representing an OR of arity $h(d)$, assuming NP $\not\subseteq$ coNP/poly. Kernel bounds where $h(d)$ is polynomially bounded in d, would therefore establish inverse polynomial lower bounds on the degree of polynomials representing an n-variable OR modulo 6. However, the current-best degree lower bound [29] is only $\Omega(\log n)$, which has not been improved in nearly two decades (cf. [4, §1.4]).

When it comes to CSPs whose constraints are of the form "the number of satisfied literals in the clause belongs to set S", many cases remain unsolved. We can prove (see Appendix A) that for constraints of the form "the number of satisfied literals is a prime number", no generalized kernel of size polynomial in n exists unless NP $\subseteq$ coNP/poly. On the other hand, Corollary 7 gives good compressions for problems of the type "the number of satisfied literals in the clause is a multiple of three". Is sparsification possible when a constraint requires the number of satisfied literals to be a square, for example?

A simple example of a CSP whose kernelization complexity is currently unclear has constraints of the form "the number of satisfied literals is one or two, modulo six". The approach of Theorem 4 fails, since there is no polynomial modulo six with root set $\{1,2\}$.

Finally, we mention that all our results extend to the setting of min-ones and max-ones CSPs, in which one has to find a satisfying assignment that sets at least, or at most, a given number of variables to true. For example, our results easily imply that Exact Hitting Set parameterized by the number of variables n has a sparsification of size $\mathcal{O}(n^2)$, which cannot be improved to $\mathcal{O}(n^{2-\varepsilon})$ unless NP $\subseteq$ coNP/poly.

References

1. David A. Mix Barrington. Some problems involving Razborov-Smolensky polynomials. In *Proceedings of the London Mathematical Society Symposium on Boolean Function Complexity*, pages 109–128. Cambridge University Press, 1992. doi:10.1017/CBO9780511526633.010.
2. David A. Mix Barrington, Richard Beigel, and Steven Rudich. Representing boolean functions as polynomials modulo composite numbers. *Computational Complexity*, 4(4):367–382, 1994. doi:10.1007/BF01263424.
3. Richard Beigel. The polynomial method in circuit complexity. In *Proc. 8th CCC*, pages 82–95, 1993. doi:10.1109/SCT.1993.336538.
4. Abhishek Bhowmick and Shachar Lovett. Nonclassical Polynomials as a Barrier to Polynomial Lower Bounds. In *Proc. 30th CCC*, volume 33 of *LIPIcs*, pages 72–87, 2015. doi:10.4230/LIPIcs.CCC.2015.72.
5. Hans L. Bodlaender. Kernelization, exponential lower bounds. In *Encyclopedia of Algorithms*. Springer, 2015. doi:10.1007/978-3-642-27848-8_521-1.
6. Hans L. Bodlaender, Bart M. P. Jansen, and Stefan Kratsch. Kernelization lower bounds by cross-composition. *SIAM J. Discrete Math.*, 28(1):277–305, 2014. doi:10.1137/120880240.
7. Hans L. Bodlaender, Stéphan Thomassé, and Anders Yeo. Kernel bounds for disjoint cycles and disjoint paths. *Theor. Comput. Sci.*, 412(35):4570–4578, 2011. doi:10.1016/j.tcs.2011.04.039.
8. Andrei A. Bulatov and Dániel Marx. Constraint satisfaction parameterized by solution size. *SIAM J. Comput.*, 43(2):573–616, 2014. doi:10.1137/120882160.
9. Marek Cygan, Fedor V. Fomin, Lukasz Kowalik, Daniel Lokshtanov, Dániel Marx, Marcin Pilipczuk, Michal Pilipczuk, and Saket Saurabh. *Parameterized Algorithms*. Springer, 2015. doi:10.1007/978-3-319-21275-3.
10. Holger Dell, Eun Jung Kim, Michael Lampis, Valia Mitsou, and Tobias Mömke. Complexity and approximability of parameterized MAX-CSPs. In *Proc. 10th IPEC*, volume 43 of *LIPIcs*, pages 294–306, 2015. doi:10.4230/LIPIcs.IPEC.2015.294.
11. Holger Dell and Dieter van Melkebeek. Satisfiability allows no nontrivial sparsification unless the polynomial-time hierarchy collapses. *J. ACM*, 61(4):23:1–23:27, 2014. doi:10.1145/2629620.
12. Michael Dom, Daniel Lokshtanov, and Saket Saurabh. Kernelization lower bounds through colors and IDs. *ACM Transactions on Algorithms*, 11(2):13, 2014. doi:10.1145/2650261.
13. Rodney G. Downey and Michael R. Fellows. *Fundamentals of Parameterized Complexity*. Texts in Computer Science. Springer, 2013.
14. J. Flum and M. Grohe. *Parameterized Complexity Theory*. Springer-Verlag, 2006.
15. Lance Fortnow and Rahul Santhanam. Infeasibility of instance compression and succinct PCPs for NP. *J. Comput. Syst. Sci.*, 77(1):91–106, 2011. doi:10.1016/j.jcss.2010.06.007.
16. Ben Green and Terence Tao. The primes contain arbitrarily long arithmetic progressions. *Annals of Mathematics*, 167(2):481–547, 2008. doi:10.4007/annals.2008.167.481.
17. Gregory Gutin. Kernelization: Constraint satisfaction problems parameterized above average. In Ming-Yang Kao, editor, *Encyclopedia of Algorithms*. Springer, 2015. doi:10.1007/978-3-642-27848-8_524-1.
18. Leslie Hogben. *Handbook of Linear Algebra, Second Edition*. Chapman and Hall/CRC, 2014.
19. Bart M. P. Jansen. On sparsification for computing treewidth. *Algorithmica*, 71(3):605–635, 2015. doi:10.1007/s00453-014-9924-2.

20 Bart M. P. Jansen. Constrained bipartite vertex cover: The easy kernel is essentially tight. In *Proc. 33rd STACS*, volume 47 of *LIPIcs*, pages 45:1–45:13, 2016. doi:10.4230/LIPIcs.STACS.2016.45.

21 Bart M. P. Jansen and Astrid Pieterse. Sparsification upper and lower bounds for graphs problems and not-all-equal SAT. In *Proc. 10th IPEC*, volume 43 of *LIPIcs*, pages 163–174, 2015. doi:10.4230/LIPIcs.IPEC.2015.163.

22 R. M. Karp. Reducibility Among Combinatorial Problems. In *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.

23 Stefan Kratsch. Recent developments in kernelization: A survey. *Bulletin of the EATCS*, 113:58–97, 2014.

24 Stefan Kratsch, Dániel Marx, and Magnus Wahlström. Parameterized complexity and kernelizability of max ones and exact ones problems. *TOCT*, 8(1):1, 2016. doi:10.1145/2858787.

25 Stefan Kratsch and Magnus Wahlström. Preprocessing of min ones problems: A dichotomy. In *Proc. 37th ICALP*, volume 6198 of *Lecture Notes in Computer Science*, pages 653–665, 2010. doi:10.1007/978-3-642-14165-2_55.

26 Daniel Lokshtanov, Neeldhara Misra, and Saket Saurabh. Kernelization - preprocessing with a guarantee. In *The Multivariate Algorithmic Revolution and Beyond - Essays Dedicated to Michael R. Fellows on the Occasion of His 60th Birthday*, volume 7370 of *Lecture Notes in Computer Science*, pages 129–161, 2012. doi:10.1007/978-3-642-30891-8_10.

27 Lásló Lovász. Chromatic number of hypergraphs and linear algebra. In *Studia Scientiarum Mathematicarum Hungarica 11*, pages 113–114, 1976.

28 Thomas J. Schaefer. The complexity of satisfiability problems. In *Proc. 10th ACM Symposium on Theory of Computing*, pages 216–226, 1978. doi:10.1145/800133.804350.

29 Gábor Tardos and David A. Mix Barrington. A lower bound on the mod 6 degree of the OR function. *Computational Complexity*, 7(2):99–108, 1998. doi:10.1007/PL00001597.

A Prime SAT

To prove the theorem, we need a result by Dell and van Melkebeek. They proved a stronger version of the following theorem in [11]. It is rephrased here to match the used definitions.

▶ **Theorem 20** ([11, Theorem 1]). *Let $d \geq 3$ be an integer. d-CNF-SAT does not have a generalized kernel of size $\mathcal{O}(n^{d-\varepsilon})$ for any $\varepsilon > 0$, unless* NP $\subseteq$ coNP/poly.

Using Theorem 20, we prove the following.

▶ **Theorem 21.** *Let* Prime-Sat *be a variant of* Generalized d-Sat *where $S := \{p \in \mathbb{N} \mid p$ is prime$\}$ for each clause.* Prime-Sat *parameterized by the number of variables does not have a polynomial kernel unless* NP $\subseteq$ coNP/poly.

Proof. We show the non-existence of a polynomial kernel by giving a linear parameter transformation from d-CNF-SAT for any d, which establishes claimed lower bound by Theorem 20. Let an instance $\mathcal{F}$ of d-CNF-SAT be given. We create an instance $\mathcal{F}'$ for Prime-Sat. Consider a clause $(\ell_1, \ldots, \ell_d)$ in $\mathcal{F}$.

It is proven in [16] that the primes contain arbitrarily long arithmetic progressions. We start by showing that each such progression has a finite length. Suppose this progression is given by $\{a + i \cdot b \mid i \in \mathbb{N} \wedge i < d\}$ for some constants $a, b, d \in \mathbb{N}$. Now note that $a + b > 1$ divides $a + (a + b + 1) \cdot b = (a + b)(b + 1)$, which bounds the length of this progression. Therefore we assume that a is chosen in such a way that $a + i \cdot d$ is not a prime.

Clause

$$(\underbrace{1, \ldots, 1}_{a \text{ copies}}, \underbrace{\neg\ell_1, \ldots, \neg\ell_1}_{b \text{ copies}}, \ldots, \underbrace{\neg\ell_d, \ldots, \neg\ell_d}_{b \text{ copies}})$$

is added to $\mathcal{F}'$. It is easy to verify that the number of satisfied literals in this clause is prime, if and only if at least one of the literals in $(\ell_1 \vee \ldots \vee \ell_d)$ is satisfied.

This clause uses multiple occurrences of the same variable and the constant 1, which is not allowed in the definition of Generalized d-Sat. This can be solved by replacing the constants by a new variables $T_1, \ldots, T_a$. These can be forced to *true* by adding clauses (T_i, T_{i+1}) for $i \in [a - 1]$, since two is a prime number while zero and one are not. For each variable x we add b distinct copies $x_1, \ldots, x_b$ and require them to be equal with clauses $(T_1, T_2, T_3, T_4, x_i, x_{i+1})$ for $i \in [b-1]$. This construction eliminates repeated variables. As the number of variables increases by a constant, this yields a valid linear-parameter transformation. ◀