Approximating minimum representations of key Horn functions

Kristóf Bérczi,[1] **Endre Boros,**[2] **Ondřej Čepek,**[3]*
Petr Kučera,[3] **Kazuhisa Makino**[4]
[1]MTA-ELTE Egerváry Research Group, Department of Operations Research, Eötvös Loránd University, Budapest, Hungary, berkri@cs.elte.hu
[2]MSIS Department and RUTCOR, Rutgers University, New Jersey, USA, endre.boros@rutgers.edu
[3]Charles University, Faculty of Mathematics and Physics, Department of Theoretical Computer Science and Mathematical Logic, Praha, Czech Republic, {cepek,kucerap}@ktiml.mff.cuni.cz
[4]Research Institute for Mathematical Sciences (RIMS), Kyoto University, Kyoto, Japan, makino@kurims.kyoto.ac.jp

Abstract

Horn functions form a subclass of Boolean functions and appear in many different areas of computer science and mathematics as a general tool to describe implications and dependencies. Finding minimum sized representations for such functions with respect to most commonly used measures is a computationally hard problem that remains hard even for the important subclass of key Horn functions. In this paper we provide logarithmic factor approximation algorithms for key Horn functions with respect to all measures studied in the literature for which the problem is known to be hard.

1 Introduction

A Boolean function of n variables is a mapping from $\{0,1\}^n$ to $\{0,1\}$. Boolean functions naturally appear in many areas of mathematics and computer science and constitute a principal concept in complexity theory. In this paper we shall study an important problem connected to Boolean functions, the so called Boolean minimization problem, which aims at finding a shortest possible representation of a given Boolean function. The formal statement of the Boolean minimization problem (BM) of course depends on (i) how the input function is represented, (ii) how it is represented on the output, and (iii) the way how the output size is measured. One of the most common representations of Boolean functions are conjunctive normal forms (CNFs), the conjunctions of clauses which are elementary disjunctions of literals. There are two usual ways how to measure the size of a CNF: the number of clauses and the total number of literals (sum of clause lengths). BM is known to be computationally very hard for both measures. It was shown in (Umans 2001) that the minimum equivalent DNF problem is Σ_2^p-complete, while a $O(n^{1-\varepsilon})$-inapproximability result was given in (Umans 1999).

Horn functions form a subclass of Boolean functions which plays a fundamental role in constructive logic and computational logic. They are important in automated theorem proving and relational databases. An important feature of Horn functions is that SAT is solvable for this class in linear time (Dowling and Gallier 1984). A CNF is Horn if every clause in it contains at most one positive literal, and it is pure Horn (or definite Horn in some literature) if every clause in it contains exactly one positive literal. A Boolean function is (pure) Horn, if it admits a (pure) Horn CNF representation. Pure Horn functions represent a very interesting concept which was studied in many areas of computer science and mathematics under several different names. The same concept appears as directed hypergraphs in graph theory and combinatorics, as implicational systems in artificial intelligence and database theory, and as lattices and closure systems in algebra and concept lattice analysis (Caspard and Monjardet 2003). Consider a pure Horn CNF $\Phi = (\overline{a} \vee b) \wedge (\overline{b} \vee a) \wedge (\overline{a} \vee \overline{c} \vee d) \wedge (\overline{a} \vee \overline{c} \vee e)$ on variables a, b, c, d, e, where $\overline{a}$ stands for the negation of a, etc. The equivalent directed hypergraph is $\mathcal{H} = (V, \mathcal{E})$ with vertex set $V = \{a, b, c, d, e\}$ and directed hyperarcs $\mathcal{E} = \{(\{a\}, b), (\{b\}, a), (\{a, c\}, d), (\{a, c\}, e)\}$. This latter can be expressed more concisely using a generalization of adjacency lists for ordinary digraphs in which all hyperarcs with the same body (also called source) are grouped together $\{a\} : b, \{b\} : a, \{a, c\} : d, e$, or can be represented as an implicational (closure) system on variables a, b, c, d, e defined by rules $a \to b, b \to a, ac \to de$.

Interestingly, in each of these areas the problem similar to BM, i.e. a problem of finding the shortest equivalent representation of the input data (CNF, directed hypergraph, set of rules) was studied. For example, such a representation can be used to reduce the size of knowledge bases in expert systems, thus improving the performance of the system. The above examples show that a "natural" way how to measure the size of the representation depends on the area. Six different measures and corresponding concepts of minimality were considered in (Ausiello, D'Atri, and Sacca 1986; Crama and Hammer 2011): (B) number of bodies, (BA) body area, (TA) total area, (C) number of clauses, (BC) num-

*Corresponding author.

ber of bodies and clauses, and (L) number of literals. For precise definitions, see Section 2. With a slight abuse of notations we shall use (B), (BA), (TA), (C), (BC) and (L) to denote both the measures and the corresponding minimization problems. The only one of these six minimization problems for which a polynomial time procedure exists to derive a minimum representation is (B). The first such algorithm appeared in database theory literature (Maier 1980). Different algorithms for the same task were then independently discovered in hypergraph theory (Ausiello, D'Atri, and Sacca 1986), and in the theory of closure systems (Guigues and Duquenne 1986).

For the remaining five measures it is NP-hard to find the shortest representation. There is an extensive literature on the intractability results in various contexts for these minimization problems (Ausiello, D'Atri, and Sacca 1986; Hammer and Kogan 1993; Maier 1980). It was shown that (C) and (L) stay NP-hard even when the inputs are limited to cubic (bodies of size at most two) pure Horn CNFs (Boros, Čepek, and Kučera 2013), and the same result extends to the remaining three measures. Note that if all bodies are of size one then the above problems become equivalent with the transitive reduction of directed graphs, which is tractable (Aho, Garey, and Ullman 1972). It should be noted that there exists many other tractable subclasses, such as acyclic and quasi-acyclic pure Horn CNFs (Hammer and Kogan 1995), and CQ Horn CNFs (Boros et al. 2009). There are also few heuristic minimization algorithms for pure Horn CNFs (Boros, Čepek, and Kogan 1998). It was shown that (C) and (L) are not only hard to solve exactly but even hard to approximate. More precisely, (Bhattacharya et al. 2010) shows that these problems are inapproximable within a factor $2^{\log^{1-\varepsilon}(n)}$ assuming $NP \subsetneq DTIME(n^{polylog(n)})$, where n denotes the number of variables. In addition, (Boros and Gruber 2014) shows that they are inapproximable within a factor $2^{\log^{1-o(1)} n}$ assuming $P \subsetneq NP$ even when the input is restricted to 3-CNFs with $O(n^{1+\varepsilon})$ clauses, for some small $\varepsilon > 0$. It is not difficult to see that the same proof extends to (BC) and (TA) as well. On the positive side, (C), (BC), (BA), and (TA) admit $(n-1)$-approximations and (L) has an $\binom{n}{2}$-approximation (Hammer and Kogan 1993). To the best of our knowledge, no better approximations are known even for pure Horn 3-CNFs.

Given a relational database, a key is a set of attributes with the property that a value assignment to this set uniquely determines the values of all other attributes (Maier 1983; Ullman 1984). Analogously, we say that a pure Horn function is *key Horn* if any of its bodies implies all other variables, that is, setting all variables in any of its bodies to one forces all other variables to one. This is a weaker concept than a database key, where setting the attributes in a key to any set of values determines the values of all remaining attributes. Key Horn functions are a generalization of a well studied class of *hydra functions* considered in (Sloan, Stasi, and Turán 2017). For this special class defined by the additional requirement that all bodies are of size two, a 2-approximation algorithm for (C) was presented in (Sloan, Stasi, and Turán 2017) while the NP-hardness for (C) was proved in (Kučera 2017). The latter result implies NP-hardness for hydra functions also for (BC), (TA), and (L). It is also easy to see that (B) and (BA) are trivial in this case.

In this paper we consider the minimization problems for key Horn functions. Any irredundant representation of a key Horn function has the same set of bodies, implying that problems (B) and (BA) are in P. We show that a simple algorithm gives a 2-approximation for (TA) and a k-approximation for (C), (BC), and (L), where k is the size of a largest body. Our paper contains two main results. The first one improves the $(n-1)$-approximation bound for (C) and (BC) to $\min\{\lceil \log n \rceil + 1, \lceil \log k \rceil + 2\}$ in the case of key Horn functions. The second result improves the $\binom{n}{2}$-approximation bound for (L) to $\frac{108}{17}\lceil \log k \rceil + 2$. Table 1 summarizes the state of the art of Horn minimization and the results presented in this paper for key Horn functions.

The structure of our paper is as follows: Section 2 introduces the necessary definitions and notation, Section 3 provides lower bounds for the measures we introduced, while Section 4 contains our results about approximation algorithms. Due to space limitations, some of the proofs are moved to the Appendix, while the NP-hardness of finding a literal minimum representation is showed in the full version of the paper (Bérczi et al. 2018).

2 Preliminaries

Let V denote a set of variables. Members of V are called *positive* while their negations are called *negative literals*. Throughout the paper, the number of variables is denoted by n. A *Boolean function* is a mapping $f : \{0,1\}^V \to \{0,1\}$. The *characteristic vector* of a set Z is denoted by χ_Z, that is, $\chi_Z(v) = 1$ if $v \in Z$ and 0 otherwise. We say that a set $Z \subseteq V$ is a *true set* of f if $f(\chi_Z) = 1$, and a *false set* otherwise.

For a subset $\emptyset \neq B \subseteq V$ and $v \in V \setminus B$ we write $B \to v$ to denote the pure Horn clause $C = v \vee \bigvee_{u \in B} \overline{u}$. Here B and v are called the *body* and *head* of the clause, respectively. That is, a pure Horn CNF can be associated with a directed hypergraph where every clause $B \to v$ is considered to be a directed hyperarc oriented from B to v. The *set of bodies* appearing in a CNF representation Φ is denoted by $\mathcal{B}_\Phi$. We will also use the notation $B \to H$ to denote $\bigwedge_{v \in H} B \to v$. By grouping the clauses with the same body, a pure Horn CNF $\Phi = \bigwedge_{B \in \mathcal{B}_\Phi} \bigwedge_{v \in H(B)} B \to v$ can be represented as $\bigwedge_{B \in \mathcal{B}_\Phi} B \to H(B)$. The latter representation is in a one-to-one correspondence with the adjacency list representation of the corresponding directed hypergraph. For any pure Horn function h the family of its true sets is closed under taking intersection and contains V. This implies that for any non-empty set $Z \subseteq V$ there exists a unique minimal true set containing Z. This set is called the *closure* of Z and is denoted by $F_h(Z)$. If Φ is a pure Horn CNF representation of h, then the closure $F_h(Z)$ can be computed in polynomial time by the following *forward chaining procedure*. Set $F^0_\Phi(Z) := Z$. In a general step, if $F^i_\Phi(Z)$ is a true set then we set $F_\Phi(Z) = F^i_\Phi(Z)$. Otherwise, let $A \subseteq V$ denote the set of all variables v for which there exists a clause

Measure	Horn		Key Horn
	Inapprox.	Approx.	Approx.
(B)	P[Maier 1980]		P[Maier 1980]
(BA)	1[Ausiello, D'Atri, and Sacca 1986]	$n-1$[Hammer and Kogan 1993]	**P**
(TA)	$2^{O(\log^{1-o(1)} n)}$[Boros and Gruber 2014]	$n-1$[Hammer and Kogan 1993]	**2**
(C)	$2^{O(\log^{1-o(1)} n)}$[Boros and Gruber 2014]	$n-1$[Hammer and Kogan 1993]	$\mathbf{\min\{\lceil \log n \rceil + 1, \lceil \log k \rceil + 2, k\}}$
(BC)	$2^{O(\log^{1-o(1)} n)}$[Boros and Gruber 2014]	$n-1$[Hammer and Kogan 1993]	$\mathbf{\min\{\lceil \log n \rceil + 1, \lceil \log k \rceil + 2, k\}}$
(L)	$2^{O(\log^{1-o(1)} n)}$[Boros and Gruber 2014]	$\binom{n}{2}$[Hammer and Kogan 1993]	$\mathbf{\min\{\frac{108}{17}\lceil \log k \rceil + 2, k\}}$

Table 1: Complexity landscape of Horn and key Horn minimization. Bold letters represent the results obtained in this paper. Here n and k respectively denote the number of variables and the size of a largest body. All problems except those labeled by P are NP-hard. Inapproximability bounds for Horn minimization hold even when the size of the bodies are bounded by k (≥ 2).

$B \to v$ of Φ with $B \subseteq F^i_\Phi(Z)$ and $v \notin F^i_\Phi(Z)$, and set $F^{i+1}_\Phi(Z) := F^i_\Phi(Z) \cup A$. The result $F_\Phi(Z)$ does not depend on the particular choice of the representation Φ, but only on the underlying function h, that is, $F_\Phi(Z) = F_h(Z)$.

A pure Horn function h is *key Horn* if it has a CNF representation of the form $\bigwedge_{B\in\mathcal{B}} B \to (V \setminus B)$ for some $\mathcal{B} \subseteq 2^V \setminus \{V\}$. We shall refer to h as $h_\mathcal{B}$. Assume now that Φ is a pure Horn CNF of the form $\bigwedge_{i=1}^m B_i \to H_i$ where $B_i \neq B_j$ for $i \neq j$. Note that the number of clauses in the CNF is $c_\Phi = \sum_{i=1}^m |H_i|$. The size of the formula can be measured in different ways:

- **(B) number of bodies**: $|\Phi|_B := m$,
- **(BA) body area**: $|\Phi|_{BA} := \sum_{i=1}^m |B_i|$,
- **(TA) total area**: $|\Phi|_{TA} := \sum_{i=1}^m (|B_i| + |H_i|)$,
- **(C) number of clauses (i.e., hyperarcs)**: $|\Phi|_C := c_\Phi$,
- **(BC) number of bodies and clauses**: $|\Phi|_{BC} := m + c_\Phi = \sum_{i=1}^m (|H_i| + 1)$,
- **(L) number of literals**: $|\Phi|_L := \sum_{i=1}^m \big((|B_i|+1)\cdot|H_i|\big)$.

These measures come up naturally in connection with directed hypergraphs, implicational systems, and CNF representations. For example,(L) corresponds to the size of a CNF when encoded in DIMACS format, a format that is widely accepted as the standard format for boolean formulas in CNF. The number of clauses (C) is an important parameter for SAT solvers when the Horn formula in question encodes a constraint which is part of a larger problem. Similarly, (TA) is the space needed to store an adjacency list of the corresponding hypergraph, and might be an important parameter for an efficient implementations. The Horn minimization problem is to find a representation that is equivalent to a given Horn formula and has minimum size with respect to $|\cdot|_*$ where $*$ denotes one of the aforementioned functions.

3 Lower bounds for the size of optimal solutions

The present section provides some simple reductions of the problem and lower bounds for the size of an optimal solution. For a family $\mathcal{B} \subseteq 2^V \setminus \{V\}$, we denote by $\mathcal{B}^\perp$ the family of minimal elements of $\mathcal{B}$. Recall that $h_\mathcal{B}$ denotes the function defined by

$$\Psi_\mathcal{B} = \bigwedge_{B\in\mathcal{B}} B \to (V \setminus B). \tag{1}$$

Lemma 1. *For any measure $(*)$ and for any $\mathcal{B} \subseteq 2^V \setminus \{V\}$, there exists a $|\cdot|_*$-minimum representation of $h_\mathcal{B}$ that uses exactly the bodies in $\mathcal{B}^\perp$.*

Proof. Take a $|\cdot|_*$-minimum representation Φ for which $|\mathcal{B}_\Phi \setminus \mathcal{B}^\perp|$ is as small as possible. First we show $\mathcal{B}_\Phi \subseteq \mathcal{B}^\perp$. Assume that $B \in \mathcal{B}_\Phi \setminus \mathcal{B}^\perp$. As B is a false set of $h_\mathcal{B}$, there must be a clause $B' \to v$ in $\Psi_\mathcal{B}$ that is falsified by χ_B, implying that $B' \subseteq B$. Therefore there exists a $B'' \in \mathcal{B}^\perp$ such that $B'' \subseteq B' \subseteq B$. If we substitute every clause $B \to v$ of Φ by $B'' \to v$, then we get another representation of $h_\mathcal{B}$ since $B'' \to v$ is a clause of $\Psi_\mathcal{B}$. Meanwhile, the $|\cdot|_*$ size of the representation does not increase while $|\mathcal{B}_\Phi \setminus \mathcal{B}^\perp|$ decreases, contradicting the choice of Φ.

Next we prove $\mathcal{B}_\Phi \supseteq \mathcal{B}^\perp$. If there exists a $B \in \mathcal{B}^\perp \setminus \mathcal{B}_\Phi$, then B is a true set of Φ while it is a false set of $h_\mathcal{B}$, contradicting the fact that Φ is a representation of $h_\mathcal{B}$. □

Recall that a *Sperner family* is family of subsets of a finite set in which none of the sets contains another. Lemma 1 has two implications. It suffices to consider Sperner families of bodies defining key Horn functions as an input, and more importantly, it is enough to consider CNFs using bodies from the input Sperner family when searching for minimum representations. For non-key Horn functions, this is not the case. For example, the function defined by implications $ab \to cd$, $abcd \to e$ has exactly two false sets, namely $\{a, b\}$ and $\{a, b, c, d\}$, and both of these sets have to appear as bodies in any representation of the function, although one of them contains the other.

From now on we assume that $\mathcal{B}$ is a Sperner family. We also assume that $\bigcup_{B\in\mathcal{B}} B = V$ and $\bigcap_{B\in\mathcal{B}} B = \emptyset$. Indeed, if a variable $v \in V \setminus \bigcup_{B\in\mathcal{B}} B$ is not covered by the bodies, then there must be a clause with head v and body in $\mathcal{B}$ in any minimum representation of $h_{\mathcal{B}}$, and actually one such clause suffices. Furthermore, if $v \in \bigcap_{B\in\mathcal{B}} B$, then we can reduce the problem by deleting it. None of these reductions affects the approximability of the problem. Recall that the size of the ground set is denoted by $|V| = n$, while $|\mathcal{B}| = m$. The size of an optimal solution with respect to measure function $|\cdot|_*$ is denoted by $OPT_*(\mathcal{B})$. Using these notations Lemma 1 implies $OPT_B(\mathcal{B}) = m$ and $OPT_{BA}(\mathcal{B}) = \sum_{B\in\mathcal{B}} |B|$. Therefore the minimization problems (B) and (BA) are solvable in polynomial time. For the remaining measures we prove the following simple lower bound.

Lemma 2. *$OPT_*(\mathcal{B}) \geq m$ for all measures $*$, and $OPT_*(\mathcal{B}) \geq n$ for $* \in \{TA, C, BC, L\}$. Furthermore, $OPT_L(\mathcal{B}) \geq \max\{n(\delta+1), 2m\}$, where δ is the size of a smallest body in $\mathcal{B}$.*

For a pair $S, T \subseteq V$ of sets, let $price_*(S,T)$ denote the minimum $|\cdot|_*$-size of a pure Horn CNF Φ for which $\mathcal{B}_\Phi \subseteq \mathcal{B}$ and $T \subseteq F_\Phi(S)$, that is,

$$price_*(S,T) = \min_{\Phi} \big\{ |\Phi|_* \mid \mathcal{B}_\Phi \subseteq \mathcal{B}, T \subseteq F_\Phi(S) \big\}. \quad (2)$$

The following lemma plays a key role in our approximability proofs.

Lemma 3. *Let $\mathcal{B} = \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_q$ be a partition of $\mathcal{B}$ and let $B_i \in \mathcal{B}_i$ for $i = 1, \ldots, q$. Then $OPT_*(\mathcal{B}) \geq \sum_{i=1}^{q} \min\{price_*(B_i, B) \mid B \in \mathcal{B} \setminus \mathcal{B}_i\}$ for all six measures $*$.*

Proof. Take a minimum representation Φ with respect to $|\cdot|_*$ which uses bodies only from $\mathcal{B}$. Such a representation exists by Lemma 1. We claim that the contribution of the clauses with bodies in $\mathcal{B}_i$ to the total size of Φ is at least $\min\{price_*(B_i, B) \mid B \in \mathcal{B} \setminus \mathcal{B}_i\}$ for each $i = 1, \ldots, q$. This would prove the lemma as the $\mathcal{B}_i$'s form a partition of $\mathcal{B}$.

To see the claim, take an index $i \in \{1, \ldots, q\}$ and let B' be the first body (more precisely, one of the first bodies) not contained in $\mathcal{B}_i$ that is reached by the forward chaining procedure from B_i with respect to Φ. Every clause that is used to reach B' from B_i has its body in $\mathcal{B}_i$ and their contribution to the size of the representation is lower bounded by $price_*(B_i, B')$, thus concluding the proof. □

4 Approximability results for (TA), (C), (BC), and (L)

Given a Sperner family $\mathcal{B} \subseteq 2^V \setminus \{V\}$, we can associate with it a complete directed graph $D_{\mathcal{B}}$ by defining $V(D_{\mathcal{B}}) = \mathcal{B}$ and $E(D_{\mathcal{B}}) = \mathcal{B} \times \mathcal{B}$. We refer to $D_{\mathcal{B}}$ as the *body graph* of $\mathcal{B}$. For any subset $E' \subseteq E(D_{\mathcal{B}})$, define

$$\Phi_{E'} = \bigwedge_{(B,B')\in E'} B \to (B' \setminus B). \quad (3)$$

Note that if $E' \subseteq E(D_{\mathcal{B}})$ forms a strongly connected spanning subgraph of $D_{\mathcal{B}}$, then $\Phi_{E'}$ is a representation of $h_{\mathcal{B}}$.

Lemma 4. *If E' is a Hamiltonian cycle in $D_{\mathcal{B}}$, then $\Phi_{E'}$ defined in* (3) *provides a k-approximation for all measures, where k is an upper bound on the sizes of bodies in $\mathcal{B}$.*

In fact, for (B) and (BA) (3) gives an optimal representation for any strongly connected spanning E'. Furthermore, if E' is a Hamiltonian cycle, we get a 2-approximation for (TA) based on the fact that the total area of any representation is lower bounded by $\sum_{B\in\mathcal{B}} |B|$.

Theorem 1. *If E' is a Hamiltonian cycle in $D_{\mathcal{B}}$, then $\Phi_{E'}$ defined in* (3) *provides a* 2*-approximation for (TA).*

Proof. $|\Phi_{E'}|_{TA} = \sum_{i=1}^{m} (|B_i| + |B_{i+1} \setminus B_i|) \leq 2\sum_{i=1}^{m} |B_i| \leq 2OPT_{TA}(\mathcal{B})$. □

The observation that a strongly connected subgraph of the body graph corresponds to a representation of $h_{\mathcal{B}}$, as in (3), suggests the reduction of our problem to the problem of finding a minimum weight strongly connected spanning subgraph (MWSCS) in a directed graph with arc-weight $price_*(B, B')$ for $(B, B') \in E(D_{\mathcal{B}})$. The optimum solution to this problem is an upper bound for the minimum $|\cdot|_*$-size of a representation of $h_{\mathcal{B}}$. As there are efficient constant-factor approximations for MWSCS (Frederickson and Jájá 1981), this approach may look promising. There are however two difficulties: for measure (L), we show that computing $price_L$ is NP-complete (Bérczi et al. 2018), and even when it is efficiently computable (for measures (C) and (BC)), the upper bound obtained in this way may be off by a factor of $\Omega(n)$ from the optimum (see (Bérczi et al. 2018) for a construction).

In what follows, we overcome these difficulties. An *in-arborescence* is a directed, rooted tree in which all edges point towards the root. An in-arborescence is called *spanning* if the underlying tree is spanning. A *branching* is a directed forest in which every connected component forms an in-arborescence. For (C), instead of a strongly connected spanning subgraph, we compute a minimum weight spanning in-arborescence and extend that to a representation of $h_{\mathcal{B}}$. The same approach works for (BC) as well. For (L), the situation is more complicated. First, we develop an efficient approximation algorithm for $price_L$. Next, we compute a minimum weight spanning in-arborescence where its root is pre-specified. Finally, we extend the corresponding CNF to a representation of $h_{\mathcal{B}}$. We show that the cost of the arborescences built is at most a multiple of the optimum by a logarithmic factor, which in turn ensures the improved approximation factor.

4.1 Clause and body-clause minimum representations

In this section we consider (C) and (BC) and show that the simple algorithm described in Procedure 1 provides the stated approximation factor. We note that a minimum weight spanning in-arborescence of a directed graph can be found in polynomial time, see (Chu 1965; Edmonds 1967).

Lemma 5. *First we observe that $price_C(B, B') = |B' \setminus B|$ for $B, B' \in \mathcal{B}$. Next, let $\overline{T}$ denote a minimum $price_C$-weight spanning in-arborescence in $D_{\mathcal{B}}$. Then $|\Phi_{\overline{T}}|_C \leq$*

Procedure 1: Approximation of (C) and (BC)

1 Determine a minimum $price_C$-weight spanning in-arborescence $\overline{T}$ of $D_{\mathcal{B}}$.
/* Denote by B_0 the body corresponding to the root of $\overline{T}$. */
2 Output $\Phi = \Phi_{\overline{T}} \wedge (B_0 \to (V \setminus B_0))$.
/* Here $\Phi_{\overline{T}}$ is defined as in (3). */

$\lceil \log k \rceil OPT_C(\mathcal{B}) + \max\{0, m-k\}$*, where* k *is an upper bound on the sizes of bodies in* $\mathcal{B}$*.*

Proof. We construct a subgraph T of $D_{\mathcal{B}}$ such that (i) it is a spanning in-arborescence, and (ii) $|\Phi_T|_C \leq \lceil \log k \rceil OPT_C(\mathcal{B}) + \max\{0, m-k\}$. This proves the lemma as the weight of T upper bounds the weight of $\overline{T}$.

We start with the digraph T_1 on node set $\mathcal{B}$ that has no arcs. In a general step of the algorithm, T_i will denote the graph constructed so far. We maintain the property that T_i is a branching, that is, a collection of node-disjoint in-arborescences spanning all nodes. In an iteration, for each such in-arborescence we choose an arc of minimum weight with respect to $price_C$ that goes from the root of the in-arborescence to some other component. We add these arcs to T_i, and for each directed cycle created, we delete one of its arcs. This results in a graph T_{i+1} with at most half the number of weakly connected components that T_i has, all being in-arborescences. We repeat this until the number of components becomes at most $\max\{1, m/k\}$. To reach this, we need at most $\lceil \log k \rceil$ iterations. Finally, we choose one of the roots of the components and add an arc from all the other roots to this one, obtaining a spanning in-arborescence T.

It remains to show that T also satisfies (ii). In the final stage, we add at most $\max\{1, m/k\} - 1$ arcs to T, which corresponds to at most $k(\max\{1, m/k\} - 1) \leq \max\{0, m-k\}$ clauses in Φ_T. Now we bound the rest of Φ_T. In iteration i, components of T_i define a partition $\mathcal{B} = \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_q$. Let us denote by B_j the body corresponding to the root of the arborescence with node-set $\mathcal{B}_j$. Let us consider the arcs $\{(B_j, B'_j) \mid j = 1, \ldots, q\}$ chosen to be added in the ith iteration. Now we obtain

$$\begin{aligned} |\Phi_{T_{i+1} \setminus T_i}|_C &\leq \sum_{j=1}^{q} price_C(B_j, B'_j) \\ &= \sum_{j=1}^{q} \min_{B \in \mathcal{B} \setminus \mathcal{B}_j} price_C(B_j, B) \\ &\leq OPT_C(\mathcal{B}). \end{aligned}$$

The first inequality follows from the construction of T. The equality follows from the criterion to choose the arcs to be added. The last inequality follows from Lemma 3. Since we have at most $\lceil \log k \rceil$ iterations, the lemma follows. □

Theorem 2. *For key Horn functions, there exists a polynomial time* $\min\{\lceil \log n \rceil + 1, \lceil \log k \rceil + 2, k\}$*-approximation algorithm for (C) and (BC), where* k *is an upper bound on the sizes of bodies in* $\mathcal{B}$*.*

Proof. We first show that Φ provided by Procedure 1 is a $\min\{\lceil \log n \rceil + 1, \lceil \log k \rceil + 2\}$-approximation for (C) and (BC). Note that Φ is a subformula of $\Psi_{\mathcal{B}}$ defined by (1) since all bodies in Φ are from $\mathcal{B}$. Furthermore, by our construction, $F_\Phi(B) = V$ for all $B \in \mathcal{B}$. This implies that the output Φ represents $h_{\mathcal{B}}$. Using Lemma 5 and the fact that we added $|V \setminus B_0| \leq n$ clauses to Φ_T in Step 2, we obtain $|\Phi|_C \leq \lceil \log k \rceil OPT_C(\mathcal{B}) + \max\{0, m-k\} + n$. By Lemma 2, this gives a $(\lceil \log k \rceil + 2)$-approximation, while setting $k = n$ gives a $(\lceil \log n \rceil + 1)$-approximation. By Lemma 1, $OPT_{BC}(\mathcal{B}) = |\mathcal{B}| + OPT_C(\mathcal{B})$. Since $|\Phi|_{BC} = |\mathcal{B}| + |\Phi|_C$, the same approximation ratios as above follow for (BC) as well.

Finally, Lemma 4 provides a different CNF that is a k-approximation for (C) and (BC). □

4.2 Literal minimum representations

In this section we consider (L). The first difficulty that we have to overcome is that, unlike in the case of (C) and (BC), computing $price_L$ is NP-hard (Bérczi et al. 2018). To circumvent this difficulty, we give an $O(1)$-approximation algorithm for $price_L(S, S')$ for any pair of sets $S, S' \subseteq V$. Note that if S does not contain a body $B \in \mathcal{B}$ then $price_L(S, S') = \infty$, hence we assume that this is not the case. We first analyze the structure of a pure Horn CNF Φ attaining the minimum in (2) for (L). Starting the forward chaining procedure from S with respect to Φ, let W_i denote the set of variables reached within the first i steps. That is, $S = W_0 \subsetneq W_1 \subsetneq \cdots \subsetneq W_t \supseteq S'$. We choose Φ in such a way that t is as small as possible (among those formulas minimizing (2) for (L)). Let $B_i \in \mathcal{B}$ be a smallest body contained in W_i for $i = 0, \ldots, t-1$ and set $B_t := S'$.

Proposition 1. $B_i \not\subseteq W_{i-1}$ *for* $i = 1, \ldots, t$.

Proposition 1 immediately implies that $|B_0| > |B_1| > \ldots > |B_{t-1}|$.

Proposition 2. $W_{i+1} \setminus W_i \subseteq B_{i+1}$ *for* $i = 0, \ldots, t-1$.

By Proposition 2, $W_{i+1} \setminus W_i = B_{i+1} \setminus (S \cup \bigcup_{j=1}^{i} B_j)$. Define $\Phi^{(1)} := \bigwedge_{i=0}^{t-1} B_i \to (B_{i+1} \setminus (S \cup \bigcup_{j=1}^{i} B_j))$. Observe that $\Phi^{(1)}$ has a simple structure which is based on a linear order of bodies $B_0, \ldots, B_t$.

Proposition 3. $|\Phi^{(1)}|_L = |\Phi|_L$.

The proposition implies that $\Phi^{(1)}$ also realizes $price_L(S, S')$. We know no efficient algorithms to compute $\Phi^{(1)}$, thus, using the next two propositions, we define a CNF that approximates $\Phi^{(1)}$ well and can be computed efficiently.

Let $i_0 = 0$ and for $j > 0$ let i_j denote the smallest index for which $|B_{i_j}| \leq |B_{i_{j-1}}|/2$. Let $r-1$ be the largest value for which $B_{i_{r-1}}$ exists and set $B_{i_r} := S'$. Now define $\Phi^{(2)} := \bigwedge_{j=0}^{r-1} B_{i_j} \to (B_{i_{j+1}} \setminus (S \cup \bigcup_{\ell=1}^{j} B_{i_\ell}))$. It is easy to see that $F_{\Phi^{(2)}}(S) \supseteq S'$.

Proposition 4. $|\Phi^{(2)}|_L \leq 2|\Phi^{(1)}|_L$.

Although $\Phi^{(2)}$ gives a 2-approximation for $|\Phi|_L$, it is not clear how we could find such a representation. Define $\Phi^{(3)} := \bigwedge_{j=0}^{r-1} B_{i_j} \to (B_{i_{j+1}} \setminus (S \cup B_{i_j}))$. The only difference between $\Phi^{(2)}$ and $\Phi^{(3)}$ is that we add unnecessary clauses to the representation. However, the next claim shows that the size of the formula cannot increase a lot.

Proposition 5. $|\Phi^{(3)}|_L \leq \frac{27}{17}|\Phi^{(2)}|_L$.

By Propositions 3, 4 and 5,

$$|\Phi^{(3)}|_L \leq \frac{27}{17}|\Phi^{(2)}|_L \leq \frac{54}{17}|\Phi^{(1)}|_L = \frac{54}{17}|\Phi|_L. \quad (4)$$

Lemma 6. *There exists an efficient algorithm to construct a CNF $\Lambda(S,S')$ such that $|\Lambda(S,S')|_L \leq \frac{54}{17}\, price_L(S,S')$, $\mathcal{B}_{\Lambda(S,S')} \subseteq \mathcal{B}$, and $F_{\Lambda(S,S')}(S) \supseteq S'$.*

Proof. We consider an extension of the body graph by adding S' to $V(D_\mathcal{B})$. We also define arc-weights by setting $w(B,B') := |B' \setminus (S \cup B)|(|B|+1)$ for $B, B' \in \mathcal{B} \cup \{S'\}$. Let B_0 be a smallest body contained in S (as defined before Proposition 1). Compute a shortest path P from B_0 to S' and define

$$\Lambda(S,S') = \bigwedge_{(B,B')\in P} B \to (B' \setminus (S \cup B)). \quad (5)$$

Note that, by definition, $|\Lambda(S,S')|_L$ is the weight of the shortest path P, while $|\Phi^{(3)}|_L$ is the length of one of the paths from S to S'. By (4), $|\Lambda(S,S')|_L \leq |\Phi^{(3)}|_L \leq \frac{54}{17}|\Phi|_L$. That is, $\Lambda(S,S')$ provides a $\frac{54}{17}$-approximation for $price_L(S,S')$ as required, finishing the proof of the lemma. □

We prove that the algorithm described in Procedure 2 provides the stated approximated factor for (L). We note that a minimum weight spanning in-arborescence of a directed graph rooted at a fixed node can be found in polynomial time, see (Chu 1965; Edmonds 1967). Let $B_{\min}$ be a smallest body in $\mathcal{B}$, let $\delta := |B_{\min}|$, and denote $\mathcal{B}' = \mathcal{B} \setminus \{B_{\min}\}$. We define the weight of an arc (B,B') in the body graph to be $w(B,B') = |\Lambda(B,B')|_L$ for $(B,B') \in E(D_\mathcal{B})$.

Procedure 2: Approximation of (L)

1 Let $B_{\min}$ be a smallest body in $\mathcal{B}$.
2 Set $w(B,B') = |\Lambda(B,B')|_L$ for $(B,B') \in E(D_\mathcal{B})$.
3 Determine a minimum w-weight spanning in-arborescence $\overline{T}$ of $D_\mathcal{B}$ such that $\overline{T}$ is rooted at $B_{\min}$.
4 Output $\Phi = \bigwedge_{(B,B')\in\overline{T}} \Lambda(B,B') \wedge (B_{\min} \to (V \setminus B_{\min}))$.
/* Here $\Lambda(B,B')$ is defined as in (5). */

Lemma 7. *Let $\overline{T}$ denote a minimum w-weight spanning in-arborescence in $D_\mathcal{B}$ such that $\overline{T}$ is rooted at $B_{\min}$. Then $\left|\bigwedge_{(B,B')\in\overline{T}} \Lambda(B,B')\right|_L \leq \left(\frac{108}{17}\lceil \log k \rceil + 1\right) OPT_L(\mathcal{B})$, where k is the size of a largest body in $\mathcal{B}$.*

Proof. We construct a subgraph T of $D_\mathcal{B}$ such that (i) it is a spanning in-arborescence rooted at $B_{\min}$, and (ii) $|\bigwedge_{(B,B')\in T} \Lambda(B,B')|_L \leq (2\lceil \log k \rceil + 1) OPT_L(\mathcal{B})$. This clearly proves the lemma as the weight of T upper bounds the weight of $\overline{T}$.

We start with the directed graph T_1 on node set $\mathcal{B}$ that has no arcs. In a general step of the algorithm, T_i will denote the graph constructed so far. We maintain the property that T_i is a branching, that is, a collection of node-disjoint in-arborescences spanning all nodes. In an iteration, for each such in-arborescence we choose an arc of minimum weight with respect to w that goes from the root of the in-arborescence to some other component. We add these arcs to T_i, and for each directed cycle created, we delete one of its arcs. This results in a graph T_{i+1} with at most half the number of weakly connected components that T_i has, all being in-arborescences. We repeat this until the number of components becomes at most $\max\{1, m/k^2\}$. To reach this, we need at most $\lceil \log k^2 \rceil \leq 2\lceil \log k \rceil$ iterations. Finally, we add an arc from all the other roots to $B_{\min}$ and delete all the arcs leaving $B_{\min}$, obtaining a spanning in-arborescence T rooted at $B_{\min}$.

It remains to show that T also satisfies (ii). In the final stage, we add at most $\max\{1, m/k^2\}$ arcs to T whose total weight is upper bounded by $(k+1)\delta \max\{1, m/k^2\} \leq \max\{n\delta, 2m\} \leq OPT_L(\mathcal{B})$, where the last inequality follows by Lemma 2. Now we bound the rest of $\bigwedge_{(B,B')\in T} \Lambda(B,B')$. In iteration i, components of T_i define a partition $\mathcal{B} = \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_q$. Let us denote by B_j the body corresponding to the root of the arborescence with node-set $\mathcal{B}_j$. Let us consider the arcs $\{(B_j, B_j') \mid j = 1, \ldots, q\}$ chosen to be added in the ith iteration. Now we obtain

$$\left|\bigwedge_{(B,B')\in T_{i+1}\setminus T_i} \Lambda(B,B')\right|_L = \sum_{j=1}^{q} w(B_j, B_j') =$$

$$\sum_{j=1}^{q} \min_{B\in\mathcal{B}\setminus\mathcal{B}_j} w(B_j, B) \leq \frac{54}{17} \sum_{j=1}^{q} \min_{B\in\mathcal{B}\setminus\mathcal{B}_j} price_L(B_j, B) \leq$$

$$\leq \frac{54}{17} OPT_L(\mathcal{B}),$$

where the first and second inequalities follow by Lemmas 6 and 3, respectively. Since we have at most $2\lceil \log k \rceil$ iterations, the lemma follows. □

Theorem 3. *For key Horn functions, there exists a polynomial time $\min\{\frac{108}{17}\lceil \log k \rceil + 2, k\}$-approximation algorithm for (L), where k is the size of a largest body in $\mathcal{B}$.*

Proof. We first show that Φ provided by Procedure 2 is a $(\frac{108}{17}\lceil \log k \rceil + 2)$-approximation for (L). Note that Φ is a subformula of $\Psi_\mathcal{B}$ defined by (1) since all bodies in Φ are from $\mathcal{B}$. Furthermore, by our construction, $F_\Phi(B) = V$ for all $B \in \mathcal{B}$. This implies that the output Φ represents $h_\mathcal{B}$. By Lemma 2, we add at most $n(\delta+1) \leq OPT_L(\mathcal{B})$ literals to $\bigwedge_{(B,B')\in T} \Lambda(B,B')$ in Step 4. This, together with Lemma 7, implies the theorem. □

References

Aho, A. V.; Garey, M. R.; and Ullman, J. D. 1972. The transitive reduction of a directed graph. *SIAM Journal on Computing* 1(2):131–137.

Ausiello, G.; D'Atri, A.; and Sacca, D. 1986. Minimal representation of directed hypergraphs. *SIAM Journal on Computing* 15(2):418–431.

Bérczi, K.; Boros, E.; Čepek, O.; Kučera, P.; and Makino, K. 2018. Approximating minimum representations of key Horn functions. *ArXiv e-prints*.

Bhattacharya, A.; DasGupta, B.; Mubayi, D.; and Turán, G. 2010. On approximate Horn formula minimization. In *International Colloquium on Automata, Languages, and Programming*, 438–450. Springer.

Boros, E., and Gruber, A. 2014. Hardness results for approximate pure Horn CNF formulae minimization. *Annals of Mathematics and Artificial Intelligence* 71(4):327–363.

Boros, E.; Čepek, O.; Kogan, A.; and Kučera, P. 2009. A subclass of Horn CNFs optimally compressible in polynomial time. *Annals of Mathematics and Artificial Intelligence* 57(3-4):249–291.

Boros, E.; Čepek, O.; and Kogan, A. 1998. Horn minimization by iterative decomposition. *Annals of Mathematics and Artificial Intelligence* 23(3-4):321–343.

Boros, E.; Čepek, O.; and Kučera, P. 2013. A decomposition method for CNF minimality proofs. *Theoretical Computer Science* 510:111–126.

Caspard, N., and Monjardet, B. 2003. The lattices of closure systems, closure operators, and implicational systems on a finite set: a survey. *Discrete Applied Mathematics* 127(2):241 – 269. Ordinal and Symbolic Data Analysis (OSDA '98), Univ. of Massachusetts, Amherst, Sept. 28-30, 1998.

Chu, Y.-J. 1965. On the shortest arborescence of a directed graph. *Scientia Sinica* 14:1396–1400.

Crama, Y., and Hammer, P. L. 2011. *Boolean functions: Theory, algorithms, and applications*. Cambridge University Press.

Dowling, W. F., and Gallier, J. H. 1984. Linear-time algorithms for testing the satisfiability of propositional horn formulae. *The Journal of Logic Programming* 1(3):267 – 284.

Edmonds, J. 1967. Optimum branchings. *Journal of Research of the National Bureau of Standards, B* 71:233–240.

Frederickson, G. N., and Jájá, J. 1981. Approximation algorithms for several graph augmentation problems. *SIAM Journal on Computing* 10(2):270–283.

Guigues, J.-L., and Duquenne, V. 1986. Familles minimales d'implications informatives résultant d'un tableau de données binaires. *Mathématiques et Sciences humaines* 95:5–18.

Hammer, P. L., and Kogan, A. 1993. Optimal compression of propositional Horn knowledge bases: complexity and approximation. *Artificial Intelligence* 64(1):131–145.

Hammer, P. L., and Kogan, A. 1995. Quasi-acyclic propositional Horn knowledge bases: optimal compression. *IEEE Transactions on knowledge and data engineering* 7(5):751–762.

Kučera, P. 2017. Hydras: Complexity on general graphs and a subclass of trees. *Theoretical Computer Science* 658:399–416.

Maier, D. 1980. Minimum covers in the relational database model. *J. ACM* 27(4):664–674.

Maier, D. 1983. *The theory of relational databases*, volume 11. Computer science press Rockville.

Sloan, R. H.; Stasi, D.; and Turán, G. 2017. Hydras: Directed hypergraphs and horn formulas. *Theor. Comput. Sci.* 658:417–428.

Ullman, J. D. 1984. *Principles of database systems*. Galgotia publications.

Umans, C. 1999. Hardness of approximating/spl sigma//sub 2//sup p/minimization problems. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, 465–474. IEEE.

Umans, C. 2001. The minimum equivalent dnf problem and shortest implicants. *Journal of Computer and System Sciences* 63(4):597–611.

Appendix

Conclusions

In this paper we study the class of key Horn functions which is a generalization of a well-studied class of hydra functions (Sloan, Stasi, and Turán 2017; Kučera 2017). Given a CNF representing a key Horn function, we are interested in finding the minimum size logically equivalent CNF, where the size of the output CNF is measured in several different ways. This problem is known to be NP-hard already for hydra CNFs for most common measures of the CNF size.

The main results of this paper are two approximation algorithms for key Horn CNFs, one for minimizing the number of clauses, and the other for minimizing the total number of literals in the output CNF. Both algorithms achieve a logarithmic approximation bound with respect to the size of the largest body in the input CNF (denoted by k). This parameter can be also defined as the size of the largest clause in the input CNF minus one. Note that k is a trivial lower bound on the number of variables (denoted by n).

These algorithms are (to the best of our knowledge) first approximation algorithms for NP-hard Horn minimization problems that guarantee a sublinear approximation bound with respect to k. It follows, that both algorithms also guarantee a sublinear approximation bound with respect to n. There are two approximation algorithms for Horn minimization known in the literature, one for general Horn CNFs (Hammer and Kogan 1993), and one for hydra CNFs (Sloan, Stasi, and Turán 2017), but both of them guarantee only a linear (or higher) approximation bound with respect to k (see Table 1 and the relevant text in the introduction section for details).

Although our analysis of Procedure 1 provides an approximation factor of $\min\{\lceil \log n \rceil + 1, \lceil \log k \rceil + 2, k\}$ for (C)

and (BC), no example is known for which the solution is tight. We believe that the proposed algorithm (possibly with slight modifications) could be used to obtain a constant factor approximation for (C) and (BC). Similarly, no example is known for which the solution provided by Procedure 2 attains a tight approximation factor. A better analysis of these procedures possibly leading to a constant factor approximation or a better lower bound than the one given in Lemma 3 is subject future research.

Acknowledgments

The research was supported by the János Bolyai Research Fellowship of the Hungarian Academy of Sciences, by the National Research, Development and Innovation Fund of Hungary under the FK_18 funding scheme (project no. NKFI-128673), by the European Union, co-financed by the European Social Fund (project no. EFOP-3.6.3-VEKOP-16-2017-00002), by Czech Science Foundation (Grant 19-19463S), and by SVV (project no. 260 453).

Proof of Lemma 2. By definition, $|\cdot|_B$ is a lower bound for all the other measures, implying $OPT_*(\mathcal{B}) \geq OPT_B(\mathcal{B}) = m$.

To see the second part, observe that $|\cdot|_C$ is a lower bound for the three other measures. Therefore it suffices to prove $OPT_C(\mathcal{B}) \geq n$. By the assumption that for every $v \in V$ there exists a $B \in \mathcal{B}$ not containing v, we can conclude by the fact that the closure $F_{h_\mathcal{B}}(B) = V$ and by the way the forward chaining procedure works that every CNF representation of $h_\mathcal{B}$ must contain at least one clause with v as its head. This implies $OPT_C(\mathcal{B}) \geq n$.

To see the last part note that every variable $v \in V$ is the head of at least one clause, the body of which is of at least size $\delta \geq 1$. Furthermore, since every body appears at least once and all clauses are of size at least 2, the claim follows. □

Proof of Lemma 4. By Lemma 1, there exists a minimum representation Φ of $h_\mathcal{B}$ such that $\mathcal{B}_\Phi = \mathcal{B}$. Since $|B' \setminus B|$ is at most k for all arcs $(B, B') \in E'$, the statement follows. □

Proof of Proposition 1. Suppose to the contrary that $B_i \subseteq W_{i-1}$ for some $1 \leq i \leq t-1$. By the definition of forward chaining, every variable $v \in W_{i+1} \setminus W_i$ is reached through a clause $B \to v$ where $B \cap (W_i \setminus W_{i-1}) \neq \emptyset$. Now substitute each such clause by $B_i \to v$. As $|B_i| \leq |B|$, the $|\cdot|_L$ size of the CNF does not increase. However, the number of steps in the forward chaining procedure decreases by at least one, contradicting the choice of Φ. Finally, $S' = B_t \subseteq W_{t-1}$ would contradict the minimality of t. □

Proof of Proposition 2. Let i be the smallest index that violates the condition. Take an arbitrary variable $v \in W_{i+1} \setminus W_i$. Then v is reached in the $(i+1)$th step of the forward chaining procedure from a body of size at least $|B_i|$. If we substitute this clause by $B_{i+1} \to v$, the resulting CNF still satisfies $F_\Phi(B_0) \supseteq S'$ but has smaller $|\cdot|_L$ size by $|B_{i+1}| < |B_i|$, contradicting the minimality of Φ. □

Proof of Proposition 3. Take an arbitrary variable $v \in B_{i+1} \setminus (S \cup \bigcup_{j=1}^{i} B_j)$ for some $i = 0, \ldots, t-1$. By the observation above, $v \in W_{i+1} \setminus W_i$. This means that Φ has at least one clause entering v, say $B \to v$, for which $B \subseteq W_i$ and so $|B| \geq |B_i|$. However, $\Phi^{(1)}$ has exactly one clause entering v, namely $B_i \to v$. This implies that $|\Phi^{(1)}|_L \leq |\Phi|_L$, and equality holds by the minimality of Φ. □

Proof of Proposition 4. Take an arbitrary variable $v \in B_{i_{j+1}} \setminus (S \cup \bigcup_{\ell=1}^{j} B_{i_\ell})$ for some $j = 0, \ldots, r-1$. Then both $\Phi^{(1)}$ and $\Phi^{(2)}$ contain a single clause entering v. Namely, v is reached from $B_{i_{j+1}-1}$ in $\Phi^{(1)}$ and from B_{i_j} in $\Phi^{(2)}$. By the definition of the sequence $i_0, i_1, \ldots, i_{r-1}$, we get $|B_{i_j}| \leq 2|B_{i_{j+1}-1}|$, concluding the proof. □

Proof of Proposition 5. Take an arbitrary variable v that appears as the head of a clause in the representation $\Phi^{(3)}$. Let j be the smallest index for which $v \in B_{i_{j+1}} \setminus (S \cup \bigcup_{\ell=1}^{j} B_{i_\ell})$. Then $\Phi^{(2)}$ contains a single clause entering v, namely $B_{i_j} \to v$. On the other hand, the set $\{B_{i_j} \to v\} \cup \{B_{i_\ell} \to v \mid \ell = j+2, \ldots, r-1\}$ contains all the clauses of $\Phi^{(3)}$ that enter v. By the definition of the sequence $i_0, i_1, \ldots, i_{r-1}$, we get $\sum_{\ell=j+2}^{r-1}(|B_{i_\ell}| + 1) = (r-j-2) + \sum_{\ell=j+2}^{r-1} |B_{i_\ell}| \leq \lfloor \log |B_{i_{j+1}}| \rfloor + |B_{i_j}|/2 - 1 \leq \lfloor \log |B_{i_j}| \rfloor + |B_{i_j}|/2 - 2$. We get at most this many extra literals in $\Phi^{(3)}$ on top of the $|B_{i_j}| + 1$ literals in $\Phi^{(2)}$. As $\lfloor \log x \rfloor/(x+1) + x/(2(x+1)) - 2/(x+1) \leq 10/17$ for $x \in \mathbb{Z}_+$, the statement follows. □