Egerváry Research Group
on Combinatorial Optimization

Technical reportS

TR-2018-12. Published by the Egerváry Research Group, Pázmány P. sétány 1/C, H–1117, Budapest, Hungary. Web site: www.cs.elte.hu/egres . ISSN 1587–4451.

Minimal representation of elementary Horn functions

Kristóf Bérczi, Endre Boros, Ondřej Čepek,
Petr Kučera, and Kazuhisa Makino

July 2018

Minimal representation of elementary Horn functions

Kristóf Bérczi*, Endre Boros†, Ondřej Čepek‡, Petr Kučera§, and Kazuhisa Makino¶

Abstract

Horn functions form a computationally tractable subclass of Boolean functions and appear in many different areas of computer science and mathematics as a general tool to describe implications and dependencies. The problem of finding a minimum representation of a Horn function is interesting both from a theoretical and a practical viewpoint. We give approximation algorithms for the problem in a special class of Horn functions.

1 Introduction

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 point** of f if $f(\chi_Z) = 1$, and a **false point** otherwise. For short, we will also use the terms **true set** and **false set**, respectively. The **sets of true** and **false sets** of f are denoted by $\mathcal{T}_f$ and $\mathcal{F}_f$.

Any Boolean function f can be represented by a **conjunctive normal form** (CNF). A CNF $\Phi = (V, \mathcal{C})$ is a conjunction of its **clauses** in $\mathcal{C}$, where each clause is a disjunction of literals. The **number of literals** in a clause C is denoted by $|C|$. A clause is **Horn** if at most one of its literals is positive, and is **pure Horn** (or **definite Horn**) if it contains exactly one positive literal. The CNF function Φ is **pure Horn**

*MTA-ELTE Egerváry Research Group, Department of Operations Research, Eötvös University, Budapest, Hungary. E-mail: berkri@cs.elte.hu.

†MSIS Department and RUTCOR, Rutgers University, New Jersey, USA. E-mail: endre.boros@rutgers.edu.

‡Charles University, Faculty of Mathematics and Physics, Department of Theoretical Computer Science and Mathematical Logic, Praha, Czech Republic. E-mail: cepek@ktiml.mff.cuni.cz.

§Charles University, Faculty of Mathematics and Physics, Department of Theoretical Computer Science and Mathematical Logic, Praha, Czech Republic. E-mail: kucerap@ktiml.mff.cuni.cz.

¶Research Institute for Mathematical Sciences (RIMS) Kyoto University, Kyoto, Japan. E-mail: makino@kurims.kyoto.ac.jp.

if all of its clauses are pure Horn, and a Boolean function f is called **pure Horn** if it has a pure Horn CNF representation.

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 hyperedge oriented from B to v. Hence we will refer to clauses and variables also as hyperedges and nodes, respectively. The **sets of bodies** and **heads** appearing in a CNF representation Φ is denoted by $\mathcal{B}_\Phi$ and $\mathcal{H}_\Phi$, respectively. 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 appearing in a pure CNF Φ we get $\Phi = \bigwedge_{i=1}^{t} B_i \to H_i$ where $t = |\mathcal{B}_\Phi|$.

For any pure Horn function h the set of its true sets, $\mathcal{T}_h$, is closed under taking intersection and contains V. This implies that for any non-empty set $Z \subseteq V$ there exists a unique smallest true set containing Z. This set is called the **forward chaining closure** of Z and is denoted by $F_h(Z)$. If Φ is a pure Horn CNF representation of h, then the forward chaining closure can be obtained by the following procedure. Set $F^0_\Phi(Z) := Z$. In a general step, if $F^i_\Phi(Z)$ is a true set then $F_h(Z) = F^i_\Phi(Z)$. Otherwise, let $A \subseteq V$ denote the set of all variables v for which there exists a clause $B \to v$ of Φ with $B \subseteq F^i_\Phi(Z)$ and $v \notin F^i_\Phi(Z)$, and set $F^{i+1}_\Phi := F^i_\Phi(Z) \cup A$. The result of the process does not depend on the particular choice of the representation Φ, but only on the underlying function h.

We call a pure Horn function h **elementary** if $F_h(B) = V$ for every inclusionwise minimal false set $B \in \mathcal{F}_h$. Note that such a function has a unique irredundant representation Φ, and $\mathcal{B}_\Phi$ is a Sperner family containing exactly the inclusionwise minimal false sets. Vice versa, an arbitrary Sperner family $\mathcal{B} \subseteq 2^V \setminus \{V\}$ can be associated with the irredundant pure Horn CNF

$$\Phi = \bigwedge_{B \in \mathcal{B}} B \to (V \setminus B),$$

which in turn represents an elementary pure Horn function, denoted by $h_\mathcal{B}$.

Given a Sperner family $\mathcal{B} \subseteq 2^V \setminus \{V\}$, we can associate with it a directed tournament $D_\mathcal{B}$ by defining $V(D_\mathcal{B}) = \mathcal{B}$, $E(D_\mathcal{B}) = \mathcal{B} \times \mathcal{B}$. We refer to $D_\mathcal{B}$ as the **body graph** of $\mathcal{B}$. If $F \subseteq E(D_\mathcal{B})$ forms a strongly connected subgraph of $D_\mathcal{B}$, then

$$\Phi_F = \bigwedge_{(B,B') \in F} B \to (B' \setminus B)$$

is a representation of $h_\mathcal{B}$.

For a pure Horn CNF $\Phi = (V, \mathcal{C})$ and a Sperner family $\mathcal{B} \subseteq 2^V \setminus \{V\}$, the **body graph of Φ with respect to $\mathcal{B}$** is a directed graph $D^\Phi_\mathcal{B}$ where $V(D^\Phi_\mathcal{B}) = \mathcal{B}$ and there is a directed edge (B, B') in $E(D^\Phi_\mathcal{B})$ if and only if $(B \to v) \in \mathcal{C}$ for each $v \in B' \setminus B$. When $\mathcal{B}$ is clear from the context, we simply call this directed graph the body graph of Φ.

Assume now that $\Phi = (V, \mathcal{C})$ is a pure Horn CNF of form $\Phi = \bigwedge_{i=1}^{t} B_i \to H_i$ where $B_i \neq B_j$ for $i \neq j$. The size of the formula can be measured in different ways:

- **number of bodies**, denoted by $|\Phi|_B := |\mathcal{B}_\Phi| = t$,
- **number of edges**, denoted by $|\Phi|_E := |\mathcal{C}| = \sum_{i=1}^{t} |H_i|$,
- **number of bodies and edges**, denoted by $|\Phi|_{BE} := |\mathcal{B}_\Phi| + |\mathcal{C}| = \sum_{i=1}^{t} (|H_i|+1)$,
- **body area**, denoted by $|\Phi|_{BA} := \sum_{i=1}^{t} |B_i|$,
- **total area**, denoted by $|\Phi|_A := \sum_{i=1}^{t} (|B_i| + |H_i|)$,
- **number of literals**, denoted by $|\Phi|_L := \sum_{C \in \mathcal{C}} |C| = \sum_{i=1}^{t} \big((|B_i| + 1) \cdot |H_i|\big)$.

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. Such a representation can be used to reduce the size of the knowledge base in a propositional expert system, thus improving the performance of the system.

Previous work. Unfortunately, it is NP-hard to find an optimal representation for almost all of these size functions (see [1]). The sole exception is the case of body minimal representations, for which polynomial time algorithms were independently discovered [1, 7, 9]. In [3], Boros et al. provided a min-max result on the minimum number of bodies appearing in the representation of a Horn function, thus giving an explanation why this case is so different from the others in terms of tractability.

In contrast, edge minimal representations are not only hard to find but even hard to approximate. Bhattacharya et al. [2] showed that the edge minimal representation problem is inapproximable within a factor $2^{O(\log^{(1-\varepsilon)}(n)}$ assuming $NP \subsetneq DTIME(n^{polylog(n)})$, while Boros and Gruber showed that it is inapproximable within a factor $2^{O(log^{1-o(1)}n)}$ assuming $P \subsetneq NP$, where n denotes the number of variables.

Contribution. The present work aims at giving approximation algorithms for the Horn minimization problem in the special class of elementary Horn functions. As mentioned earlier, the body minimal representation problem is well studied and can be solved in polynomial time. Hence we concentrate on the remaining size definitions.

2 Preliminaries

The problem that we consider is the following: we are given a Sperner family $\mathcal{B} \subseteq 2^V \setminus \{V\}$, and the aim is to find a minimum pure Horn CNF representation of $h_\mathcal{B}$. 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 $OPT_*(\mathcal{B})$.

We may assume w.l.o.g. that $\bigcup_{B \in \mathcal{B}} B = V$, since if there is a node $v \in V \setminus \bigcup_{B \in \mathcal{B}} B$ not covered by any of the bodies, then in any minimal representation of $h_\mathcal{B}$ there must be a clause with head v and body in $\mathcal{B}$, and actually one such clause suffices. We may also assume that $\bigcap_{B \in \mathcal{B}} B = \emptyset$, since if there exists a node $v \in V$ which is contained in all members of $\mathcal{B}$ then we can reduce the problem by simply deleting it.

We start with easy lower bounds on the size of an optimal solution.

Lemma 2.1. $OPT_*(\mathcal{B}) \geq \max\{m, n\}$ *for* $* \in \{E, BE, A, L\}$.

Proof. Observe that all measure functions are lower bounded by $|\cdot|_E$ except $|\cdot|_B$ and $|\cdot|_{BA}$, hence it suffices to prove the statement for $* = E$.

If we start the forward chaining from a body $B \in \mathcal{B}$, we have to reach all the other nodes as the function is elementary. This means that, in any representation, there must be at least one hyperedge with body B for every $B \in \mathcal{B}$, implying $OPT_E(\mathcal{B}) \geq m$.

On the other hand, every node $v \in V$ has to be reachable by forward chaining from every $B \in \mathcal{B}$. By the assumption that no node is contained in all members of $\mathcal{B}$, every representation must contain at least one hyperedge with head node v, implying $OPT_E(\mathcal{B}) \geq n$. □

The following two lemmas play a key role in our approximation algorithms.

Lemma 2.2. *There exists a* $|\cdot|_*$*-minimal representation that only uses bodies from* $\mathcal{B}$.

Proof. Take a $|\cdot|_*$-minimal representation Φ for which $|\mathcal{B}_\Phi \setminus \mathcal{B}|$ is as small as possible. If $\mathcal{B}_\Phi \setminus \mathcal{B} = \emptyset$ then we are done, hence assume that this is not the case. Let $B \in \mathcal{B}_\Phi \setminus \mathcal{B}$. As B is a false set, there must be a body $B' \in \mathcal{B}$ s.t. $B' \subset B$. If we substitute every edge $B \to v$ of Φ by $B' \to v$, then the $|\cdot|_*$ size of the representation does not increase while $|\mathcal{B}_\Phi \setminus \mathcal{B}|$ decreases, contradicting the choice of Φ. □

For a body $B \in \mathcal{B}$ and set $S \subseteq V$ let $price_*(B, S)$ denote the minimum $|\cdot|_*$**-cost of reaching** S **from** B **by forward chaining using only bodies in** $\mathcal{B}$, that is,

$$price_*(B, S) = \min\left\{|\Phi|_* \;\middle|\; \mathcal{B}_\Phi \subseteq \mathcal{B}, F_\Phi(B) \supseteq S\right\}. \tag{1}$$

For a given CNF Φ, we denote by $\Phi_*(B, S)$ the minimum $|\cdot|_*$-sized sub-CNF of Φ for which S is reachable from B by forward chaining, that is,

$$\Phi_*(B, S) = \arg\min\left\{|\Phi'|_* \;\middle|\; \Phi' \subseteq \Phi, F_{\Phi'}(B) \supseteq S\right\}. \tag{2}$$

Lemma 2.3. *Let* $\mathcal{B} = \mathcal{B}_1 \cup \ldots \mathcal{B}_q$ *be a partition of* $\mathcal{B}$ *and let* $B_i \in \mathcal{B}_i$ *for* $i = 1, \ldots, q$. *Then*

$$OPT_*(\mathcal{B}) \geq \sum_{i=1}^{q} \min_{B \notin \mathcal{B}_i} price_*(B_i, B) \tag{3}$$

for $* \in \{B, E, BE, BA, A, L\}$.

Proof. Take a minimal representation Φ with respect to $|\cdot|_*$ which uses bodies only from $\mathcal{B}$. Such a representation exists by Lemma 2.2. We claim that the contribution of the edges with bodies in $\mathcal{B}_i$ to the total size of the solution is at least $\min_{B \notin \mathcal{B}_i} price_*(B_i, B)$ for each $i = 1, \ldots, q$. This would prove the lemma as the $\mathcal{B}_i$'s form a partition of $\mathcal{B}$.

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 the earliest when starting the forward chaining from B_i. Every edge 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. □

3 Edge minimal representations

As before, let $\mathcal{B} \subseteq 2^V \setminus \{V\}$ be a Sperner family corresponding to an elementary pure Horn function. Recall that $\bigcup_{B\in\mathcal{B}} B = V$ and $\bigcap_{B\in\mathcal{B}} B = \emptyset$. Now we are ready to prove our first result.

Theorem 3.1. *For elementary pure Horn functions, there exists an efficient* $(\lceil \log n \rceil + 1)$*-approximation algorithm for the edge minimal representation problem.*

Proof. The high level idea of the algorithm is as follows. In a first phase, we will construct a CNF so that its body graph with respect to $\mathcal{B}$ contains a branching consisting of a 'few' components. We start from the empty formula Φ_0 and the empty digraph F_0. In a general step of the algorithm, Φ_i will denote the CNF constructed so far and F_i will denote a branching of $D_{\mathcal{B}}^{\Phi_i}$. Then Φ_{i+1} is determined in such a way that for each component -which is an arborescence- of F_i, there exists an arc in $E(D_{\mathcal{B}}^{\Phi_{i+1}})$ leaving it. This results in a branching F_{i+1} having at most half that many components as F_i.

When the number of components in the branching becomes small enough, we add all the possible hyperedge leaving the root bodies of the components in a second phase. This step makes the body graph of the formula strongly connected, which means that we get a representation of the pure Horn function.

Now we give a detailed description of the algorithm. As mentioned before, Φ_0 denotes the empty formula and F_0 is the empty digraph on node-set $\mathcal{B}$. Now we show how to define Φ_{i+1} and F_{i+1}. As F_i is a branching, it is the node-disjoint collection of arborescences $A_1, \ldots, A_q$ (where an arborescence may consist of a single node). The node-sets of these arborescences defines a partition of $\mathcal{B}$ into subsets $\mathcal{B}_1 \cup \ldots \cup \mathcal{B}_q$ where $\mathcal{B}_j$ is the set of bodies corresponding to the node-set of A_j. Now define $B_j \in \mathcal{B}_j$ as the body corresponding to the root-node of A_j. For each $j = 1, \ldots, q$, let B'_j be a body attaining the minimum in $\min_{B\notin\mathcal{B}_j} |B \setminus B_j|$ and set $\Phi_{i+1} := \Phi_i \wedge (\bigwedge_{j=1}^q B_j \to B'_j)$.

Claim 3.2. $price_E(B_j, B) = |B \setminus B_j|$.

Proof. Take a CNF attaining the minimum in (1). As every node in $B \setminus B_j$ is reached by forward chaining starting from B_j, each such node must be a head of at least one hyperedge. That is, the CNF contains at least $|B \setminus B_j|$ hyperedges. However, $B_j \to (B \setminus B_j)$ uses exactly $|B \setminus B_j|$ edges, hence $price_E(B_j, B) = |B \setminus B_j|$ as required. □

By Lemma 2.3 and Claim 3.2, the total number of edges added is $\sum_{j=1}^q \min_{B\notin\mathcal{B}_j} |B \setminus B_j| = \sum_{j=1}^q \min_{B\notin\mathcal{B}_j} price_E(B_j, B) \leq OPT_E(\mathcal{B})$.

Now there is an arc from B_j to B'_j in $E(D_{\mathcal{B}}^{\Phi_{i+1}})$ for each $j = 1, \ldots, q$. If we add these arcs to F_i simultaneously, we get a directed graph which is almost a branching, except that every component contains exactly one directed cycle. We can break these cycles by deleting a newly added arc in all of them, thus obtaining a branching F_{i+1} consisting of at most $q/2$ arborescences.

After $\lceil \log n \rceil$ extension steps, the number of components of the branching decreases to at most $m/2^{\lceil \log n \rceil} \leq m/n$ and the first phase ends. Let $A_1, \ldots, A_q$ denote the arborescences forming the connected components of $F_{\lceil \log n \rceil}$. As before, the root body

of A_j is denoted by B_j. Now set $\Phi = \Phi_{\lceil \log n \rceil} \wedge (\bigwedge_{j=1}^{q} B_j \to (V \setminus B_j))$. This step ensures that the body graph $D_{\mathcal{B}}^{\Phi}$ contains all complete out-stars whose center node is among the B_j's. These out-stars together with the branching $F_{\lceil \log n \rceil}$ form a strongly-connected digraph, hence Φ is a representation of the Horn function in question.

By Lemma 2.1, the number of edges added in the second phase is bounded by $m/n \cdot n = m \leq OPT_E(\mathcal{B})$. Hence the total size of our solution is bounded by $(\lceil \log n \rceil + 1) \cdot OPT_E(\mathcal{B})$ as stated. □

The approach used in the proof of Theorem 3.1 improves the previously known best approximation factor for $(k+1)$-CNF's [10]. Note that this corresponds to the case when $|B| \leq k$ for every $B \in \mathcal{B}$.

Theorem 3.3. *For elementary pure Horn functions with $|B| \leq k$ for every $B \in \mathcal{B}$, there exists an efficient $(\lceil \log k \rceil + 2)$-approximation algorithm for the edge minimal representation problem.*

Proof. Perform the same steps as in the proof of Theorem 3.1, but now the first phase stops after $\lceil \log k \rceil$ steps. The number of components in the branching thus obtained is at most m/k. Let $A_1, \ldots, A_q$ denote the arborescences forming the connected components of $F_{\lceil \log k \rceil}$ with root bodies $B_1 \ldots, B_q$. Fix an arbitrary body $B_0 \in \mathcal{B}$ and set $\Phi = \Phi_{\lceil \log k \rceil} \wedge (\bigwedge_{j=1}^{q} B_j \to (B_0 \setminus B_j)) \wedge (B_0 \to (V \setminus B_0))$. This step ensures that the body graph $D_{\mathcal{B}}^{\Phi}$ contains an arc from B_j to B_0 for $j = 1, \ldots, q$, and also arcs from B_0 to any other body B. These arcs together with the branching $F_{\lceil \log k \rceil}$ form a strongly-connected digraph, hence Φ is a representation of the Horn function in question.

By Lemma 2.1, the number of edges added in the second phase is bounded by $m/k \cdot k + n = m + n \leq OPT_E(\mathcal{B}) + OPT_E(\mathcal{B})$. Hence the total size of our solution is bounded by $(\lceil \log k \rceil + 2) \cdot OPT_E(\mathcal{B})$. □

4 Body-and-edge minimal, body area minimal and total area minimal representations

In this section we discuss minimal representations of elementary pure Horn functions when the size of the representation is measured with respect to $|\cdot|_{BE}$, $|\cdot|_{BA}$ or $|\cdot|_{A}$. Observe that each member of $\mathcal{B}$ must appear as a body of an edge in any representation, hence the body area of any representation is lower bounded by $\sum_{B \in \mathcal{B}} |B|$. In fact, finding a body area minimal representation is trivial.

Theorem 4.1. *For elementary pure Horn functions, a body area minimal representation can be found in polynomial time.*

Proof. The CNF defined as $\Phi = \bigwedge_{B \in \mathcal{B}} B \to (V \setminus B)$ is a valid representation of the pure Horn function in question and only uses bodies from $\mathcal{B}$. □

Also, the number of bodies in any representation is lower bounded by $|\mathcal{B}|$, which together with Theorem 3.1 imply the following.

Theorem 4.2. *For elementary pure Horn functions, there exists an efficient* $(\lceil \log n \rceil + 1)$*-approximation algorithm for the body-and-edge minimal representation problem.*

Proof. Clearly, $|\cdot|_{BE} = |\cdot|_B + |\cdot|_E$. By the observation above, $|\cdot|_B| \geq |\mathcal{B}|$. The algorithm presented in Theorem 3.1 provides a representation Φ which uses bodies only from $\mathcal{B}$. That is, $|\Phi|_{BE} = |\Phi|_B + |\Phi|_E \leq OPT_B(\mathcal{B}) + (\lceil \log n \rceil + 1) \cdot OPT_E(\mathcal{B}) \leq (\lceil \log n \rceil + 1) \cdot OPT_{BE}(\mathcal{B})$. □

Based on the fact that the body area of any representation is lower bounded by $\sum_{B \in \mathcal{B}} |B|$, the total area minimal representation can be approximated within a constant factor.

Theorem 4.3. *For elementary pure Horn functions, there exists an efficient* 2*-approximation algorithm for the total area minimal representation problem.*

Proof. Take an arbitrary ordering $B_1, \ldots, B_t$ of the bodies in $\mathcal{B}$ and set $\Phi = \bigwedge_{i=1}^{t} B_i \to (B_{i+1} \setminus B_i)$ where $B_{t+1} = B_1$. The size of Φ is $|\Phi|_A = \sum_{i=1}^{t} (|B_i| + |B_{i+1} \setminus B_i|) \leq 2 \cdot \sum_{i=1}^{t} |B_i| \leq 2 \cdot OPT_A(\mathcal{B})$. □

5 Literal minimal representations

In this section we give an approximation algorithm for the literal minimal representation problem. The main idea of the algorithm is similar to the the proof of Theorem 3.1. Starting from the empty CNF Φ_0, we extend the CNF step-by-step. Meanwhile, we keep tracking of a branching in its body graph. The first phase stops when the number of components in the branching becomes small enough. Then, in a second phase, we add all the possible hyperedges leaving the root bodies of these components, thus making the body graph of the formula strongly connected.

The main challenge is to bound the number of edges that are added in the first phase. In the edge minimal case, Claim 3.2 ensured that the $price_E(B_i, B)$ values could be computed and so Lemma 2.3 was applicable. However, for literal minimal representations even determining $price_L(B_i, B)$ is NP-complete. To circumvent this difficulty, first we give a 4-approximation algorithm for determining $price_L(B_i, B)$.

Lemma 5.1. *Let* $B_0 \in \mathcal{B}$ *be a body and* $S \subseteq V$ *be a set. There exists an efficient* 4*-approximation algorithm for determining* $price_L(B_0, S)$*.*

Proof. Assume that Φ is a CNF attaining the minimum in (1) which provides the shortest forward chaining in time. Starting the forward chaining procedure from B_0, let W_i denote the set of nodes reached until time i. That is, $B_0 = W_0 \subset W_1 \subset \ldots \subset W_t \supseteq S$. Choose $B_i \in \mathcal{B}$ to be the smallest body in W_i for $i = 0, \ldots, t-1$ and set $B_t := S$.

Claim 5.2. $B_i \not\subseteq W_{i-1}$ *for* $i = 1, \ldots, t$*.*

Proof. The bodies in $\mathcal{B}$ form a Sperner system, hence B_0 is the single body contained in W_0, implying $B_1 \not\subseteq W_0$. Also, $S = B_t \not\subseteq W_{t-1}$ as otherwise the CNF contains unnecessary clauses, contradicting its minimality.

Suppose to the contrary that $B_i \subseteq W_{i-1}$ for some $2 \leq i \leq t-1$. By the definition of forward chaining, every node $v \in W_{i+1} \setminus W_i$ is reached through an edge $B \to v$ where $B \cap (W_i \setminus W_{i-1}) \neq \emptyset$. Now substitute each such edge by $B_i \to v$. As $|B_i| \leq |B|$, the $|\cdot|_L$ size of the CNF does not increase. However, the time length of the forward chaining decreases by at least one, contradicting the choice of Φ. □

Claim 5.2 immediately implies that $|B_0| > |B_1| > \ldots > |B_{t-1}|$.

Claim 5.3. *$W_{i+1} \setminus W_i \subseteq B_{i+1}$ for $i = 0, \ldots, t-1$.*

Proof. Let i be the smallest index that violates the condition. Take an arbitrary node $v \in W_{i+1} \setminus W_i$. 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 edge 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. □

By Claim 5.3, $W_{i+1} \setminus W_i = B_{i+1} \setminus (\bigcup_{j=0}^{i} B_j)$. Define $\Phi' := \bigwedge_{i=0}^{t-1} B_i \to (B_{i+1} \setminus \bigcup_{j=0}^{i} B_j)$.

Claim 5.4. *$|\Phi'|_L = |\Phi|_L$.*

Proof. Take an arbitrary node $v \in B_{i+1} \setminus (\bigcup_{j=0}^{i} B_j)$ 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 edge entering v, say $B \to v$, for which $B \subseteq W_i$ and so $|B| \geq |B_i|$. However, Φ' has exactly one edge entering v, namely $B_i \to v$. This implies that $|\Phi'|_L \leq |\Phi|_L$, and equality holds by the minimality of Φ. □

Let i_j denote the smallest index for which $|B_{i_j}| \leq |B_0|/2^j$ and let $r-1$ be the largest value for which $B_{i_{r-1}}$ exists. Furthermore, set $B_{i_r} := S$. Now define $\Phi'' := \bigwedge_{j=0}^{r-1} B_{i_j} \to (B_{i_{j+1}} \setminus \bigcup_{\ell=0}^{j} B_{i_\ell})$.

Claim 5.5. *$|\Phi''|_L \leq 2|\Phi'|_L$.*

Proof. Take an arbitrary node $v \in B_{i_{j+1}} \setminus (\bigcup_{\ell=0}^{j} B_{i_\ell})$ for some $j = 0, \ldots, r-1$. Then both Φ' and Φ'' contain a single edge entering v. Namely, v is reached from $B_{i_{j+1}-1}$ in Φ' and from B_{i_j} in Φ''. 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 of the claim. □

Although Φ'' gives a 2-approximation of Φ, it is still not clear how we could find such a representation. Define $\Phi''' := \bigwedge_{j=0}^{r-1} B_{i_j} \to (B_{i_{j+1}} \setminus (B_{i_j} \cup B_0))$. The only difference between Φ'' and Φ''' is that we add unnecessary edges to the representation. However, the next claim shows that the size of the formula cannot increase a lot.

Claim 5.6. *$|\Phi'''|_L \leq 2|\Phi''|_L$.*

Proof. Take an arbitrary node v that appears as the head of an edge in the representation Φ'''. Let j be the smallest index for which $v \in B_{i_{j+1}} \setminus (\bigcup_{\ell=0}^{j} B_{i_\ell})$. Then Φ'' contains a single edge entering v, namely $B_{i_j} \to v$. On the other hand, the edges of Φ''' that enter v form a subset of $\{B_{i_j} \to v, \ldots, B_{i_{r-1}} \to v\}$. By the definition of the sequence $i_0, i_1, \ldots, i_{r-1}$, we get $|B_{i_j}| + \ldots + |B_{i_{r-1}}| \leq 2|B_{i_j}|$, and the claim follows. □

By Claims 5.4, 5.5 and 5.6,

$$|\Phi'''|_L \leq 2|\Phi''|_L \leq 4|\Phi'|_L = 4|\Phi|_L.$$

The reason for considering Φ''' instead of Φ is that the body graph of Φ''' with respect to $\mathcal{B} \cup \{S\}$ is a directed path (more precisely contains a directed path) from B_0 to S.

Set the weight of an arc (B, B') of the body graph of $\mathcal{B} \cup \{S\}$ to $w_{B_0}(B, B') := |B' \setminus (B \cup B_0)|$. A shortest path P from B_0 to S corresponds to a CNF

$$\Phi_P = \bigwedge_{(B,B') \in P} B \to (B' \setminus (B \cup B_0)). \tag{4}$$

Note that such a shortest path can be found in polynomial time. Now $|\Phi_P|_L \leq |\Phi'''| \leq 4|\Phi|_L$, so Φ_P provides a 4-approximation for $price_L(B_0, S)$ as required, finishing the proof of the lemma. □

Theorem 5.7. *For elementary pure Horn functions, there exists an efficient* $(8\lceil \log n \rceil + 1)$*-approximation algorithm for the body area minimal representation problem.*

Proof. We use an approach similar to the proof of Theorem 3.1. Starting from an empty CNF Φ_0, we add edges to the formula while we keep tracking of a branching in the body graph of the formula with respect to $\mathcal{B}$. The first phase stops when the number of components in the branching decreases under a given bound. Then, in a second phase, we add further edges leaving the root bodies of the branching, thus making the body graph strongly connected.

In more details, let Φ_0 denote the empty formula and F_0 be the empty digraph on node-set $\mathcal{B}$. In a general step, Φ_i is the CNF constructed so far and F_i is a branching in the body graph of Φ_i. As F_i is a branching, it is the node-disjoint collection of arborescences $A_1, \ldots, A_q$ (where an arborescence may consist of a single node). The node-sets of these arborescences defines a partition of $\mathcal{B}$ into subsets $\mathcal{B}_1 \cup \ldots \cup \mathcal{B}_q$ where $\mathcal{B}_j$ is the set of bodies corresponding to the node-set of A_j. Now define $B_j \in \mathcal{B}_j$ as the body corresponding to the root-node of A_j.

Given an index j, we have seen that determining a body B attaining the minimum in $\min_{B \notin \mathcal{B}_j} price_L(B_j, B)$ is difficult. However, by Lemma 5.1, $price_L(B_j, B)$ can be approximated within a factor of 4 for each body $B \notin \mathcal{B}_i$ by considering shortest paths in the body graph, where the weight of an arc (B, B') is set to $w_{B_i}(B, B') := |B' \setminus (B \cup B_0)|$. That is, if P_j is a shortest path in the body graph from B_j to outside $\mathcal{B}_j$ with other end node B'_j, then Φ_{P_j} defined in (4) satisfies

$$|\Phi_{P_j}|_L \leq 4 \min_{B \notin \mathcal{B}_j} price_L(B_j, B).$$

Take such a path P_j for each $j = 1, \ldots, q$, and set $\Phi_{i+1} := \Phi_i \wedge (\bigwedge_{j=1}^{q} \Phi_{P_j})$. By the above observations and Lemma 2.3, the total size of the newly added edges is

$$\sum_{j=1}^{q} |\Phi_{P_j}|_L \leq 4 \sum_{j=1}^{q} \min_{B \notin \mathcal{B}_j} price_L(B_j, B) \leq 4 \cdot OPT_L(\mathcal{B}).$$

The arcs of the paths appear in $E(D_{\mathcal{B}}^{\Phi_{i+1}})$. If we add these paths to F_i simultaneously, we get a directed graph which may contain cycles, but as every path connected at least two components of F_i, $E(D_{\mathcal{B}}^{\Phi_{i+1}})$ contains a branching consisting of at most $q/2$ arborescences. We choose such a branching to be F_{i+1}.

After $\lceil \log n^2 \rceil$ extension steps, the number of components of the branching decreases to at most $m/2^{\lceil \log n^2 \rceil} \leq m/n^2$ and the first phase ends. Let $A_1, \ldots, A_q$ denote the arborescences forming the connected components of $F_{\lceil \log n^2 \rceil}$. As before, the root body of A_j is denoted by B_j. Now set $\Phi = \Phi_{\lceil \log n^2 \rceil} \wedge (\bigwedge_{j=1}^{q} B_j \to (V \setminus B_j))$. This step ensures that the body graph $D_{\mathcal{B}}^{\Phi}$ contains all complete out-stars with center node among the B_j's. These out-stars together with the branching $F_{\lceil \log n \rceil}$ form a strongly-connected digraph, hence Φ is a representation of the Horn function in question.

By Lemma 2.1, the number of edges added in the second phase is bounded by $m/n^2 \cdot n^2 = m \leq OPT_L(\mathcal{B})$. Hence the total size of our solution is bounded by $(4\lceil \log n^2 \rceil + 1) \cdot OPT_L(\mathcal{B}) \leq (8\lceil \log n \rceil + 1) \cdot OPT_L(\mathcal{B})$ as stated. □

The proof of the following result is analogous to that of Theorem 3.3.

Theorem 5.8. *For elementary pure Horn functions with $|B| \leq k$ for every $B \in \mathcal{B}$, there exists an efficient $(8\lceil \log k \rceil + 2)$-approximation algorithm for the literal minimal representation problem.*

6 Relation to minimum weight strongly connected subgraphs

Given a strongly connected graph $D = (V, E)$ and non-negative weights $w : E \to \mathbb{Z}_+$, we denote by $\texttt{MWSCS}(D, w)$ the problem of finding a minimum weight subset $F \subseteq E$ of the arcs such that (V, F) is also strongly connected. We denote by $\texttt{mwscs}(D, w) = w(F)$ the weight of such a minimum weight arc subset. Problem MWSCS is an NP-hard problem, for which polynomial time approximation algorithms are known. For the case of uniform weights a 1.61-approximation was given by Khuller et al. [8]. For general weights a simple 2-approximation is due to Fredericson and Jájá [6]. Note that in the case of general weights, we can assume that D is a complete directed tournament.

As it was observed already in Section 1, there is a natural relation of the above problem to the minimization of an elementary pure Horn function. Let us consider a Sperner hypergraph $\mathcal{B} \subseteq 2^V \setminus \{V\}$ and the corresponding Horn function

$$h_{\mathcal{B}} = \bigwedge_{B \in \mathcal{B}} B \to (V \setminus B).$$

The body graph of $\mathcal{B}$ was a complete directed tournament $D_{\mathcal{B}}$ where $V(D_{\mathcal{B}}) = \mathcal{B}$. Define a weight function w on the arcs of this tournament by setting $w_{\mathcal{B}}(B, B') = |B' \setminus B|$ for all $B, B' \in \mathcal{B}$, $B \neq B'$. Then any solution $F \subseteq E(G_{\mathcal{B}}) = \mathcal{B} \times \mathcal{B}$ of problem $\texttt{MWSCS}(D_{\mathcal{B}}, w_{\mathcal{B}})$ defines a representation of $h_{\mathcal{B}}$:

$$\Phi(F) = \bigwedge_{(B,B') \in E(G_{\mathcal{B}})} B \to (B' \setminus B).$$

It is immediate to see that $OPT_E(\mathcal{B}) \leq w_{\mathcal{B}}(F)$ holds. Thus, it is natural to expect that a polynomial time approximation of problem $\texttt{MWSCS}(D_{\mathcal{B}}, w_{\mathcal{B}})$ provides also a good approximation for $OPT_E(\mathcal{B})$. This however turns out to be false.

As before, we use $n = |V|$ to denote the number of variables of the Horn functions we consider.

Theorem 6.1.

$$\max_{\mathcal{B}} \frac{\texttt{mwscs}(D_{\mathcal{B}}, w_{\mathcal{B}})}{OPT_E(\mathcal{B})} \geq \frac{n}{12}.$$

Proof. For the proof, let us recall some basic facts on finite projective spaces from the book [5].

The finite projective space $PG(d, q)$ of dimension d over a finite field $GF(q)$ of order q (prime power) has $n = q^d + q^{d-1} + \cdots + q + 1$ points. Subspaces of dimension k are isomorphic to $PG(k, q)$ for $0 \leq k < d$, where 0-dimension subspaces are the points themselves. The number of subspaces of dimension $k < d$ is

$$N_k(d, q) \;=\; \prod_{i=0}^{k} \frac{q^{d+1-i} - 1}{q^{i+1} - 1},$$

and the number of points of such a subspace is $q^k + q^{k-1} + \cdots + q + 1$. In particular, the number of subspaces of dimension $d - 1$ is $N_{d-1}(d, q) = n$. If F and F' are two distinct subspaces of dimension k, then

$$2k - d \leq dim(F \cap F') \leq k - 1.$$

Furthermore, any $k + 1$ points belong to at least one subspace of dimension k.

Let us also recall that $PG(d, q)$ has a cyclic automorphism. In other words the points of $PG(d, q)$ can be identified with the integers of the cyclic group $\mathbb{Z}_n$ of modulo n addition such that if $F \subseteq \mathbb{Z}_n$ is a subspace of dimension k, then $F + i = \{f + i \mod n \mid f \in F\}$ is also a subspace of dimension k and F and $F + i$ are distinct. Furthermore, if $H \subseteq \mathbb{Z}_n$ is a subspace of dimension $d - 1$ then the family $\mathcal{H} = \{H + i \mid i \in \mathbb{Z}_n\}$ contains all subspaces of $PG(d, q)$ of dimension $d - 1$. In the rest of this section we use $+$ for the modulo n addition of integers.

Lemma 6.2. *For every $k = 0, ..., d-1$ there exists a unique subspace of dimension k that contains $\{0, 1, ..., k\}$.*

Proof. By the properties we recalled above it follows that there is at least one such subspace for every $0 \leq k < d$. We prove that there is at most one by induction on k. For $k = 0$ this is obvious, since the points are the only subspaces of dimension 0. Assume next that the claim is already proved for all $k' < k$, and consider indirectly two distinct subspaces F and F' of dimension k both of which contains the set $\{0, 1, ..., k\}$. Then $F \cap F'$ and $(F - 1) \cap (F' - 1) = (F \cap F') - 1$ are two distinct subspaces of dimension $k' < k$ and both contain $\{0, 1, ..., k-1\}$, contradicting our assumption, and thus proving our claim. □

Thus, by Lemma 6.2 there exists a unique subspace $H \subseteq \mathbb{Z}_n$ of dimension $d-1$ that contains $\{0, 1, ..., d-1\}$. Let us also introduce the set $D = \{0, 1, ..., d\}$.

Lemma 6.3. *$d \notin H$.*

Proof. Assume to the contrary that $d \in H$. Then the set $\{0, 1, ..., d-1\}$ is contained by both H and $H - 1 = H + (n-2)$, contradicting Lemma 6.2, since H and $H-1$ are distinct subspaces of dimension $d-1$. □

Let us now define $\mathcal{B} := \mathcal{H} \cup \{D + i \mid i \in \mathbb{Z}_n\}$, and observe that for any distinct pair $B \in \mathcal{H}$ and $B' \in \mathcal{B}$ we have $|B \setminus B'| \geq q^{d-1}$. Since in any solution $F \subseteq \mathcal{B} \times \mathcal{B}$ we must have an arc entering B for all $B \in \mathcal{H}$, we get

$$\texttt{mwscs}(D_{\mathcal{B}}, w_{\mathcal{B}}) \;\geq\; n \cdot q^{d-1}.$$

On the other hand, we have that

$$\Phi \;=\; (D \to (\mathbb{Z}_n \setminus D)) \wedge \left(\bigwedge_{i \in \mathbb{Z}_n} (H + i) \to d + i \right) \wedge \left(\bigwedge_{i \in \mathbb{Z}_n} (D + i) \to d + 1 + i \right)$$

is a representation of $h_{\mathcal{B}}$ and $|\Phi|_E \leq 3n$. Choosing $q = 2$ and $d > 1$, we get

$$\texttt{mwscs}(G_{\mathcal{B}}, w_{\mathcal{B}}) \;\geq\; \frac{n}{12} \cdot OPT_E(\mathcal{B}),$$

completing the proof of the theorem. □

Acknowledgement

Kristóf was supported by the Hungarian National Research, Development and Innovation Office – NKFIH grant K109240 and by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

References

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

[2] A. Bhattacharya, B. DasGupta, D. Mubayi, Gy. Turán, On approximate horn formula minimization, In: *ICALP, Lecture Notes in Computer Science*, (1), volume 6198: 438–450, 2010.

[3] E. Boros, O. Čepek, K. Makino, A combinatorial min-max theorem and minimization of pure-horn functions, In *International Symposium on Artificial Intelligence and Mathematics*, 2016.

[4] H. K. Büning, T. Lettmann, Propositional Logic: Deduction and Algorithms, Cambridge University Press, 1999.

[5] P. Dembowski Finite Geometries, Springer Science and Business Media, 1968 (2012).

[6] G.N. Fredericson, J. JáJá, Approximation algorithms for several graph augmentation problemss, *SIAM Journal on Computing*, 10(2):270–283, 1981.

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

[8] S. Khuller, B. Raghavachari, N. Young, On strongly connected digraphs with bounded cycle length, *Discrete Applied Mathematics*, 69(2):281–289, 1996.

[9] D. Maier, Minimum covers in relational database model, *Journal of ACM*, 27(4):664–674, 1980.

[10] R.H. Sloan, D. Stasi, G. Turán, Hydras: Directed hypergraphs and Horn formulas, In *International Workshop on Graph-Theoretic Concepts in Computer Science*, 237–248, 2012.