Akademie věd České republiky
Ústav teorie informace a automatizace

Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation

RESEARCH REPORT

M. Studený:

LP relaxations and pruning for characteristic imsets

No. 2323 June 2012

ÚTIA AV ČR, P. O. Box 18, 182 08 Prague, Czech Republic
Telex: 122018 atom c, Fax: (+42) (2) 688 4903
E-mail: utia@utia.cas.cz

This report constitutes an unrefereed manuscript which is intended to be submitted for publication. Any opinions and conclusions expressed in this report are those of the author(s) and do not necessarily represent the views of the Institute.

LP relaxations and pruning for characteristic imsets

Milan Studený[*]
Institute of Information Theory and Automation of the ASCR
email: studeny@utia.cas.cz

June 10, 2012

Abstract

The geometric approach to learning a Bayesian network (BN) structure is based on the idea to represent every BN structure by a certain vector. Suitable such a zero-one vector representative is the *characteristic imset*, introduced in [20]. This concept allows one to re-formulate the task of finding the global maximum of a score over BN structures as an integer linear programming (ILP) problem.

In this research report, extensions of characteristic imsets are considered which additionally encode chain graphs without flags equivalent to acyclic directed graphs. The main contribution is the *LP relaxation* of the corresponding polytope, that is, a polyhedral description (= in terms of linear inequalities) of the domain of the respective ILP problem. The advantage of this approach is that, as a by-product of the ILP optimization procedure, one may directly get the *essential graph*, which is a traditional graphical BN representative. Another topic discusses in the report is the method of search space reduction from [7]; in the considered context it leads to suitable *pruning* components of the characteristic imsets, that is, to the reduction of the length of these vector representatives.

Keywords: learning Bayesian network structure; quality criterion; integer linear programming; characteristic imset; LP relaxation; essential graph; search space reduction.

1 Introduction

Learning a *Bayesian network* (BN) structure by a score-and-search method means to maximize a *quality criterion* $\mathcal{Q}$, also named a *score* or a *scoring function.* The criterion is a real function of the (acyclic directed) graph G and the observed database D. The value $\mathcal{Q}(G, D)$ evaluates how the BN structure defined by the graph G fits the database D.

Two important technical assumptions on the criterion $\mathcal{Q}$, which were pinpointed in the literature in connection with computational aspects of this maximization task, are that $\mathcal{Q}$ should be *score equivalent* [2] and (additively) *decomposable* [3]. The geometric approach is to represent every BN structure by a certain vector so that such a criterion $\mathcal{Q}$ becomes an affine function of the vector representative. This idea was introduced already in [18] and then deepened in [19]. A suitable (uniquely determined) such a zero-one vector BN representative seems to be the *characteristic imset*, introduced in [20].

[*]This research has been supported by the grant GAČR n. 201/08/0539.

Jaakkola *et al.* [10] and Cussens [4, 5] came independently with an analogous geometric approach. The main difference is that they used certain special zero-one vector codes of (acyclic) directed graphs to represent (non-uniquely) BN structures. On the other hand, they made more progress with the practical application of *integer linear programming* (ILP) tools. To overcome technical problems with the exponential length of their vectors they utilized the idea of the reduction of the search space from [6], based on a particular form of databases and criteria occurring in practice; for deeper results on this topic see [7].

In [21], both methods of BN structure vector representation were compared and it was found that the characteristic imset can be viewed as a (many-to-one) linear function of the above mentioned graph-code. Finally, Lindner [12] performed some preliminary computational experiments based on the characteristic imset approach; for an overview of this approach see [22] and [9].

In this report, an extended vector BN structure representative is introduced, which includes the characteristic imset and, moreover, encodes a certain special graph (equivalent to an acyclic directed graph). The main result is a *polyhedral characterization* of the domain of the respective ILP problem. Specifically, a set of linear inequalities is presented such that the only vectors with integer components in the polyhedron specified by those inequalities are the above mentioned extended vector representatives. It is called the *LP relaxation* (of the corresponding polytope).

The inequalities are classified in four groups. The number of inequalities in the first two groups is polynomial in the number of variables (= nodes of the graph), while the number of remaining inequalities is exponential. However, provided the length of the vector representatives is limited/reduced to a polynomial number by the idea of from [7], the number of inequalities in the third group can be reduced to a polynomial number as well. The inequalities in the fourth group correspond to acyclicity restrictions and two versions of these inequalities are presented. In general, they cannot be reduced to a polynomial number, but, owing to their natural graphical interpretation, the method of iterative constraint adding may be applied to solve the respective ILP problems.

Another advantage of this extended vector BN structure representative is that one can get, as a result of solving an ILP problem, the *essential graph*, which is known as a standard (unique) graphical BN representative [1].

2 Preliminaries

Let N be a finite non-empty set of *variables*; to avoid the trivial case, assume $|N| \geq 2$. In statistical context, the elements of N correspond to random variables in consideration; in graphical context, they correspond to nodes.

2.1 Graphical concepts

A *hybrid graph* over (the set of *nodes* N) is a graph with two types of edges between (distinct) nodes $i, j \in N$, namely directed edges, called *arrows*, and denoted like $i \to j$ (or $j \leftarrow i$), and undirected edges, called *lines*, and denoted like $i - j$ (or $j - i$). No multiple edges are allowed between two nodes of a hybrid graph H, which means, if $i \to j$ in H, then $\neg(j \to i \text{ in } H)$ and $\neg(i - j \text{ in } H)$. If there is an edge between nodes i and j, we say they are *adjacent*. A graph is *undirected* if it only has lines; it is *directed* if it only has arrows.

A *cycle* of the length $m \geq 3$ in a hybrid graph H is a sequence $\rho : i_0, i_1, \ldots, i_m = i_0$, where $i_1, \ldots, i_m$ are distinct nodes, and, for each $r = 0, \ldots, m-1$, $[i_r, i_{r+1}]$ is an edge in H. The cycle ρ is *chordless*, or minimal, if there is no other edge in H between nodes in $\{i_1, \ldots, i_m = i_0\}$ besides those which form the cycle ρ. An undirected graph is called *chordal* if it has no chordless cycle of the length $m \geq 4$.

The cycle ρ is *directed* if $i_r \rightarrow i_{r+1}$ in H for each $r = 0, \ldots, m-1$. A directed graph is *acyclic* if it has no directed cycle. An equivalent definition of an acyclic directed graph G is that there exists an ordering $b_1, \ldots, b_{|N|}$ of all nodes in N which is consistent with the direction of arrows: $b_i \rightarrow b_j$ in G implies $i < j$. The set of *parents* of a node $i \in N$ in a (directed) graph G is the set $pa_G(i) \equiv \{ j \in N;\ j \rightarrow i \text{ in } G\}$.

The cycle ρ is *semi-directed* if $i_0 \rightarrow i_1$ in H and, for each $r = 1, \ldots, m-1$ one has either $i_r \rightarrow i_{r+1}$ in H or $i_r - i_{r+1}$ in H. We say that a hybrid graph H is *3-acyclic* if it has no semi-directed cycle of the length 3.

A set of nodes $C \subseteq N$ in a hybrid graph H is *connected* if, for every pair $i, j \in C$, there exists an undirected path between i and j, that is, a sequence of (distinct) nodes $i = i_1, \ldots, i_s = j$, $s \geq 1$ such that $i_r - i_{r+1}$ in H for $r = 1, \ldots, s-1$. The maximal connected sets (with respect to inclusion) are called the *connected components* of H.

A hybrid graph H is called a *chain graph* if (all) its components can be ordered into a sequence $C_1, \ldots, C_m$, $m \geq 1$, called a *chain*, such that if $i \rightarrow j$ in H, then $i \in C_r$ and $j \in C_s$ with $r < s$. An equivalent definition of a chain graph is that it is a hybrid graph which has no semi-directed cycle, or alternatively, which has no semi-directed chordless cycle; see Lemma 2.1 in [15]. Thus, chain graphs are, in fact, acyclic hybrid graphs, they involve acyclic directed graphs. Also, clearly, every chain graph is a 3-acyclic hybrid graph.

A *flag* in a hybrid graph H is an induced subgraph (over three nodes) of the form $i \rightarrow j - k$, which means that i and k are not adjacent. An *immorality* in H is an induced subgraph (over three nodes) of the form $i \rightarrow k \leftarrow j$ (that is, i and j are not adjacent in H).

2.2 Bayesian network structure

In statistical context, each variable $i \in N$ is assigned a finite (individual) *sample space* X_i (of possible values); assume $|\mathsf{X}_i| \geq 2$ for each $i \in N$ to avoid technical problems. The *joint sample space* is the Cartesian product $\mathsf{X}_N \equiv \prod_{i \in N} \mathsf{X}_i$. For $A \subseteq N$, the *projection* of an element $x = [x_i]_{i \in N}$ of X_N to A is the configuration (= the list) of values $x_A = [x_i]_{i \in A}$. The symbol X_A will denote the respective *projection space*. Thus, for $A = \emptyset$, the projection space $\mathsf{X}_\emptyset$ is a singleton, consisting of the empty list; for $\emptyset \neq A$ one has $\mathsf{X}_A = \prod_{i \in A} \mathsf{X}_i$.

A *Bayesian network* (BN) is a pair (G, P), where G is an acyclic directed graph with the node set N and P a *Markovian* probability distribution (with respect to G) on the *joint sample space*. This means P that satisfies conditional independence restrictions determined by G; see [11] for details. The *BN structure* defined by an acyclic directed graph G is the class of Markovian probability distributions with respect to G on (fixed) $\prod_{i \in N} \mathsf{X}_i$.

However, different graph over N can be *Markov equivalent*, which means they define the same BN structure. Classic graphical characterization of equivalent graphs by Verma and Pearl [23] says they are Markov equivalent iff they have the same adjacencies and the same immoralities.

2.3 Essential graph

A standard (unique) graphical representative of a BN structure is the following.

Definition 2.1 Let $\mathcal{G}$ be a Markov equivalence class of acyclic directed graphs over N. The *essential graph* G^* of $\mathcal{G}$ is a hybrid graph over N defined as follows:

- $a \to b$ in G^* if $a \to b$ in every G from $\mathcal{G}$,
- $a — b$ in G^* if there are graphs G_1 and G_2 in $\mathcal{G}$ with $a \to b$ in G_1 and $a \leftarrow b$ in G_2.

This terminology and the first graphical characterization of essential graphs was given by Anderson, Madigan and Perlman [1]. It implies that every essential graph is a chain graph and has no flags. Actually, the class of *chain graphs without flags* appear to be important in this context. One can introduce a graphical concept of equivalence for these graphs.

Definition 2.2 Two chain graphs without flags G and H over N are *equivalent* if they have the same adjacencies and immoralities. Given two such graphs, we say that H is *larger* than G if, for any $i, j \in N$, $i \to j$ in H implies $i \to j$ in G.

The meaning of this equivalence of G and H is that they define the same statistical model; c.f. Lemma 2 in [17]. The following characterization of the essential graphs, proved as Corollary 4 in [17], will be utilized later.

Lemma 2.1 *Let $\mathcal{G}$ be a Markov equivalence class of acyclic directed graphs over N and $\mathcal{H}$ the equivalence class of chain graphs without flags such that $\mathcal{G} \subseteq \mathcal{H}$. Then G^* is the largest graph in $\mathcal{H}$.*

Further important observation is that a chain graph H without flags is equivalent to an acyclic directed graph G iff the induced subgraphs of H for its components are chordal undirected graphs; see Lemma 3 in [17].

2.4 Characteristic imset

This algebraic representative of a BN structure was introduced in [20]. For our purpose, the following equivalent definition is suitable.

Definition 2.3 Let G be an acyclic directed graph over N. The *characteristic imset* for G can be introduced as a zero-one vector c_G with components $\mathsf{c}_G(S)$ where $S \subseteq N$, $|S| \geq 2$, such that

$$\mathsf{c}_G(S) = 1 \quad \Leftrightarrow \quad \exists\, i \in S \text{ such that } S \setminus \{i\} \subseteq pa_G(i). \tag{1}$$

The point is that two acyclic directed graphs G and H over N are Markov equivalent iff $\mathsf{c}_G = \mathsf{c}_H$; see § 3 of [9]. Moreover, Corollary 2 in [9] implies that, for different $i, j, k \in N$,

(i) i and j are adjacent in G iff $\mathsf{c}_G(\{i, j\}) = 1$,

(ii) $i \to k \leftarrow j$ is an immorality in G iff $\mathsf{c}_G(ijk) = 1$ and $\mathsf{c}_G(ij) = 0$.

In particular, one can observe that the characteristic imset c_G is uniquely determined by its values $\mathsf{c}_G(S)$ for $S \subseteq N$, $2 \leq |S| \leq 3$.

It appears to be suitable to have a formula for the characteristic imset on the basis of any graph H in the class $\mathcal{H}$ from Lemma 2.1. For this purpose one needs the following auxiliary concept.

Definition 2.4 We say that a hybrid graph H over $S \subseteq N$ has a *super-terminal component* if there exists non-empty set $K \subseteq S$ such that

- K is a *complete set* in H, which means, for each pair of distinct nodes $i, k \in K$, one has $i - k$ in H,
- $\forall\, j \in S \setminus K \;\; \forall\, i \in K$ one has $j \rightarrow i$ in H.

It makes no problem see that a super-terminal component K, if exists, is uniquely determined.

The following result follows directly from Theorem 2 in [9]:

Lemma 2.2 *Let H be a chain graph without flags equivalent to an acyclic directed graph G. For any $S \subseteq N$, $|S| \geq 2$ one has $\mathsf{c}_G(S) = 1$ iff the induced subgraph of H for S (denoted by H_S) has a super-terminal component.*

2.5 Straightforward codes of graphs

Jaakkola *et al.* [10] and Cussens [4, 5] used a special method for vector encoding (acyclic) directed graphs over N. The vector η_G encoding G has components indexed by pairs $(i|B)$, where $i \in N$ and $B \subseteq N \setminus \{i\}$. Specifically, it is defined as follows:

$$\eta_G(i|B) = \begin{cases} 1 & B = pa_G(i), \\ 0 & \text{otherwise}. \end{cases} \tag{2}$$

In their paper [10], Jaakkola *et al.* mentioned special inequalities for η_G, which they called the *cluster inequalities.* These inequalities correspond to sets $S \subseteq N$, $|S| \geq 2$:

$$1 \leq \sum_{i \in S} \; \sum_{B \subseteq N \setminus S} \eta_G(i|B)\,. \tag{3}$$

The meaning of the inequality (3) is that there exists at least one node i in the set S with $pa_G(i) \cap S = \emptyset$. In other words, the induced subgraph G_S has at least one *initial node*, that is, a node $i \in S$ with no $j \in S$ such that $j \rightarrow i$ in G_S. Because G_S is an acyclic directed graph over S, the existence of such a node $i \in S$ is obvious. That's why (3) holds for every acyclic directed graph G

In [21], the relation of the characteristic imset to this straightforward code of G was established. Actually, c_G is a linear function of η_G given by

$$\mathsf{c}_G(S) = \sum_{i \in S} \; \sum_{B,\, S \setminus \{i\} \subseteq B \subseteq N \setminus \{i\}} \eta_G(i|B) \quad \text{for } S \subseteq N,\ |S| \geq 2. \tag{4}$$

Indeed, (4) follows directly from Definition 2.3: clearly, at most one node $i \in S$ with $S \setminus \{i\} \subseteq pa_G(i)$ exists in an acyclic directed graph G_S.

2.6 Learning a BN structure

Learning a BN structure means to determine it on the basis of an observed (complete) *database* D, which is a sequence $x_1, \ldots, x_d$ of elements of the joint sample space $\prod_{i\in N} \mathsf{X}_i$ ($d \geq 1$ is the length of the database). The *projection* of D to $A \subseteq N$ is the sequence of respective projections $x^1_A, \ldots, x^d_A$, denoted by the symbol D_A. Given a database D of the length $d \geq 1$ and $y \in \mathsf{X}_A$, $A \subseteq N$ we use a special notation $d_{[y]} \equiv |\{l;\ 1 \leq l \leq d,\ x^l_A = y\}|$ for the number of occurrences of y in the database D_A. In particular, for the empty list $y \in \mathsf{X}_\emptyset$, one has $d_{[y]} = d$. The concatenation of two configurations $y \in \mathsf{X}_A$ and $z \in \mathsf{X}_B$ for $A, B \subseteq N$, $A \cap B = \emptyset$ will be denoted by $[y, z]$; it belongs to $\mathsf{X}_{A\cup B}$.

Learning a BN structure is often done by maximizing some *quality criterion*, also named a *score*, which is a bivariate real function $(G, D) \mapsto \mathcal{Q}(G, D)$, where G is an acyclic directed graph over N and D a database. The value $\mathcal{Q}(G, D)$ quantitatively evaluates how the BN structure defined by the graph G is good to explain the occurrence of the database D. For formal definition of the relevant concept of *statistical consistency* see [13].

Given the observed database D, the goal is to maximize $G \mapsto \mathcal{Q}(G, D)$. Since the aim is learn a BN structure, a natural assumption is that the criterion $\mathcal{Q}$ we are going to maximize should be *score equivalent*, which means, for every database D,

$$\mathcal{Q}(G, D) = \mathcal{Q}(H, D) \quad \text{whenever } G \text{ and } H \text{ are Markov equivalent acyclic directed graphs.}$$

The crucial technical assumption is that $\mathcal{Q}$ should be additively *decomposable*, which means, it has the form

$$\mathcal{Q}(G, D) = \sum_{i\in N} q_D(i \,|\, pa_G(i)) \,, \tag{5}$$

where the summands $q_D(*|*)$ are called *local scores*. Moreover, it is required here that each local score term $q_D(i|B)$, for $i \in N$ and $B \subseteq N \setminus \{i\}$, only depends on the database projection $D_{\{i\}\cup B}$. One can re-write (5) in terms of η_G:

$$\mathcal{Q}(G, D) = \sum_{i\in N} \sum_{B\subseteq N\setminus\{i\}} q_D(i|B) \cdot \eta_G(i|B), \tag{6}$$

which allows one to interpret $\mathcal{Q}$ as (the restriction of) a linear function of η_G.

A well-known example of such a score is Schwarz's [14] *Bayesian information criterion* (BIC), whose local scores are given as follows: for $i \in N$, $B \subseteq N \setminus \{i\}$,

$$\mathsf{bic}_D(i|B) = \underbrace{\sum_{y\in \mathsf{X}_B} \sum_{z\in \mathsf{X}_i} d_{[y,z]} \cdot \ln \frac{d_{[y,z]}}{d_{[y]}}}_{\mathsf{mll}_D(i|B)} - \frac{\ln d}{2} \cdot \underbrace{\{r(i) - 1\} \cdot \prod_{j\in B} r(j)}_{\mathsf{dim}(i|B)} \,, \tag{7}$$

where $r(i) = |\mathsf{X}_i|$, $i \in N$ are the cardinalities of the individual sample spaces. In that formula, the conventions $0 \cdot \ln \frac{0}{*} \equiv 0$ and $\prod_{j\in\emptyset} r(j) \equiv 1$ are applied. The first term here is the local score of the *maximized log-likelihood criterion* (MLL); what is subtracted is a penalty term reflecting the length of the database and the local contribution to the *dimension* (DIM). The penalty term is to quantify the complexity of the statistical model given by G.

Another common example is the *Bayesian Dirichlet Equivalence* (BDE) score [8], whose local scores are given by the following formula: for $i \in N$, $B \subseteq N \setminus \{i\}$,

$$\mathsf{bde}_D(i|B) = \sum_{y \in \mathsf{X}_B} \left\{ \ln \frac{\Gamma(\alpha_{[y]})}{\Gamma(\alpha_{[y]} + d_{[y]})} - \sum_{z \in \mathsf{X}_i} \ln \frac{\Gamma(\alpha_{[y,z]})}{\Gamma(\alpha_{[y,z]} + d_{[y,z]})} \right\}, \tag{8}$$

where the α-terms are so-called *hyper-parameters*. Typically, for $y \in \mathsf{X}_B$, they are given by the formula $\alpha_{[y]} = \alpha^* \cdot (|\mathsf{X}_B|)^{-1} = \alpha^* \cdot (\prod_{j \in B} r(j))^{-1}$ with some fixed $\alpha^* > 0$, called the *equivalent sample size.*

3 What is the aim?

The final aim of the learning procedure proposed in this report is to get a zero-one vector, which is an extension of the characteristic imset and, simultaneously, encodes the essential graph. Therefore, the considered vectors have two kinds of components:

- $\mathsf{c}(S)$ for sets $S \subseteq N$, $|S| \geq 2$, intended to encode the values of the characteristic imset,
- $\mathsf{a}(i \to j)$ for ordered pairs $i, j \in N$, $i \neq j$, intended to encode the presence of an arrow $i \to j$ in a graph (= the codes of *arrowheads*).

However, it seems to be convenient and advantageous, *in the first phase*, to allow a wider class of graphs. The idea is to consider a certain class of graphs which includes both the essential graphs and acyclic directed graphs; see below § 3.1. What is presented in the report is the LP relaxation of the respective polytope (= of the convex hull of the set of considered vector codes; see § 3.2. This allows one to specify properly the domain of the respective ILP problem; see § 6.2.

However, the pure version of the obtained ILP problem is still too complex to be solved by standard ILP software packages: this os becuase of both the exponential length of vector codes and of the exponential number of inequalities (in the number of variables $|N|$). Therefore, we describe the idea of the reduction of the search space, which allows one to shorten the length of considered vectors codes substaintially; see § 7.2. This pre-processing step (for the first phase), based on a particular form of databases and criteria occuring in practice, allows us to reduce the number of inequalities as well.

Then, *in the second phase*, when the solution of the first ILP problem is already found and the characteristic imset c determined, another ILP problem is formulated. The second ILP problem is based on the same LP relaxation and its solution is the desired simultaneous code of the essential graph and the characteristic imset c. The second ILP procedure is much simpler than the first one (clearly polynomial in $|N|$); it can be viewed as an LP version of the reconstruction algorithm for the essential graph on thr basis of the characteristic imset. Morerover, the second ILP procedure can be modified in such a way that the solution will be one of the respective acyclic directed graphs; see § 8. The summary of the whole procedure, including complexity considerations, is given in § 9.

3.1 The class of graphs and encoding

The output of the first phase should be a graph over N which falls within the class of *chain graphs without flags that are equivalent to acyclic directed graphs.*

Now, we are going to introduce special zero-one vector codes for even more general graphs, because this appears to be convenient.

Definition 3.1 Let H be a hybrid graph over N which is 3-acyclic and has no flags. Then we ascribe to H a zero-one vector $(\mathsf{a}_H, \mathsf{c}_H)$ with components determined as follows:

$$\mathsf{a}_H(i \to j) = \begin{cases} 1 & i \to j \text{ in } H, \\ 0 & \text{otherwise,} \end{cases} \quad \mathsf{c}_H(S) = \begin{cases} 1 & H_S \text{ has a super-terminal component,} \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

where $i, j \in N$ are distinct and $S \subseteq N$, $|S| \geq 2$.

Observe, that in case $|S| = 2$, H_S has a super-terminal component iff the nodes in S are adjacent in H. Thus, the relation (9) means that, for distinct $i, j \in N$,

$$\mathsf{c}_H(ij) = 1 \;\Leftrightarrow\; [i,j] \text{ is an edge in } H. \tag{10}$$

In case $|S| = 3$, as H is 3-acyclic, H_S has a super-terminal component iff either H_S is *quasi-complete*, that is, every pair of distinct nodes in S is adjacent in H, or H_S is an immorality. In particular, for distinct $i, j, k \in N$,

$$\mathsf{c}_H(ijk) = 1 \;\Leftrightarrow\; \begin{cases} \text{either } [i,j], [i,k], [j,k] \text{ are all edges in } H, \\ \text{or, up to a permutation of } i,j,k, \text{ one has } i \to k \leftarrow j \text{ in } H. \end{cases} \tag{11}$$

3.2 What are the inequalites?

The inequalities are classified in four groups and none of them is superfluous (see below).

The *basic non-negativity inequalities* are as follows:

(b.1) $\forall\, i, j \in N$ distinct $\qquad 0 \leq \mathsf{a}(i \to j)$,

(b.2) $\forall\, S \subseteq N$, $|S| = 3, 4 \qquad 0 \leq \mathsf{c}(S)$.

The *consistency inequalities* mainly relate the a-part of the vector with its c-part:

(c.1) $\forall\, i, j \in N$ distinct $\qquad \mathsf{a}(i \to j) + \mathsf{a}(j \to i) \leq \mathsf{c}(ij)$,

(c.2) $\forall\, i, j \in N$ distinct $\qquad \mathsf{c}(ij) \leq 1$,

(c.3) $\forall\, i, j, k \in N$ distinct $\qquad 2 \cdot \mathsf{c}(ijk) \leq 2 \cdot \mathsf{c}(ij) + \mathsf{a}(i \to k) + \mathsf{a}(j \to k)$,

(c.4) $\forall\, i, j, k \in N$ distinct $\qquad \mathsf{a}(i \to j) + \mathsf{c}(jk) \leq 1 + \mathsf{c}(ijk) + \mathsf{a}(j \to k)$,

(c.5) $\forall\, i, j, k \in N$ distinct $\qquad \mathsf{a}(i \to j) + \mathsf{c}(jk) + \mathsf{c}(ik) \leq 2 + \mathsf{a}(i \to k) + \mathsf{a}(k \to j)$.

The *extension inequalities* ensure the c-part to be determined by $\mathsf{c}(S)$ for $2 \leq |S| \leq 3$:

(e.1) $\forall\, S \subseteq N$, $|S| \geq 3 \qquad \sum_{i \in S} \mathsf{c}(S \setminus \{i\}) \leq 2 + (|S| - 2) \cdot \mathsf{c}(S)$,

(e.2) $\forall\, S \subseteq N$, $|S| \geq 4 \qquad (|S| - 1) \cdot \mathsf{c}(S) \leq \sum_{i \in S} \mathsf{c}(S \setminus \{i\})$.

Finally, the *acyclicity inequalities* only concern the c-part and ensure that the solution is the graph in the considered class:

(a.1) $\forall\, S \subseteq N,\ |S| \geq 4 \qquad \sum_{T\subseteq S,\, |T|\geq 2} \mathsf{c}(T)\cdot(-1)^{|T|} \leq |S|-1.$

The meaning (= interpretation) of above inequalities will be explained later, after the necessity proof (see § 4.2). As concerns their number:

- the number of the basic non-negativity constraints is polynomial in $n \equiv |N|$, namely $2\cdot\binom{n}{2}+\binom{n}{3}+\binom{n}{4}$,
- the number of the consistency inequalities is also polynomial, namely $2\cdot\binom{n}{2}+15\cdot\binom{n}{3}$,
- the number of the extension inequalities is exponential, namely $2\cdot 2^n-\binom{n}{3}-2\cdot\binom{n}{2}-2\cdot n-2$, but there is a fair hope that one can in practice reduce their number polynomially (see § 9),
- the number of the acyclicity inequalities is also exponential, namely $2^n-\binom{n}{3}-\binom{n}{2}-n-1$, and one cannot reduce it to a polynomial amount in general, but can possibly design a reasonable way of avoiding the use of all of them; see § 9.

Now, examples are given, that none of the above inequalities is a consequence of the others:

(b.1) take $N=\{i,j\}$ and $\mathsf{a}(i\to j)=-1$ while $\mathsf{a}(j\to i)=\mathsf{c}(ij)=0$,

(b.2) take $N=S$ and put $\mathsf{c}(S)=-1$ while $\mathsf{a}\equiv 0$ and $\mathsf{c}(T)=0$ for $T\subset S$,

(c.1) take $N=\{i,j\}$ and $\mathsf{c}(ij)=\mathsf{a}(i\to j)=\mathsf{a}(j\to i)=1$,

(c.2) take $N=\{i,j\}$ and $\mathsf{c}(ij)=2$ while $\mathsf{a}(i\to j)=\mathsf{a}(j\to i)=1$,

(c.3) take $N=\{i,j,k\}$ and $\mathsf{c}(ijk)=\mathsf{c}(ik)=\mathsf{c}(jk)=1$ while $\mathsf{c}(ij)=0$ and $\mathsf{a}\equiv 0$,

(c.4) take $N=\{i,j,k\}$ and $\mathsf{a}(i\to j)=\mathsf{c}(jk)=1$ while $\mathsf{c}(ij)=1$ but $\mathsf{c}(ik)=\mathsf{c}(ijk)=0$ and a vanishes in other components,

(c.5) take $N=\{i,j,k\}$ and $\mathsf{a}(i\to j)=1$, $\mathsf{c}\equiv 1$ while a vanishes in other components,

(e.1) in case $|S|=3$ take $N=\{i,j,k\}$ and $\mathsf{c}(ij)=\mathsf{c}(ik)=\mathsf{c}(jk)=1$, while $\mathsf{c}(ijk)=0$ and $\mathsf{a}\equiv 0$; if $|S|\geq 4$ take $N=S$, choose distinct $i,j\in S$ and put $\mathsf{a}(i\to k)=\mathsf{a}(j\to k)=1$ for every $k\in S\setminus\{i,j\}$, a vanishing for other components, $\mathsf{c}(ij)=\mathsf{c}(S)=0$, while $\mathsf{c}(T)=1$ for $T\subset S$, $T\neq\{i,j\}$,

(e.2) take $N=S$ and $\mathsf{c}(S)=1$, while $\mathsf{c}(T)=0$ for $T\subset S$, $\mathsf{a}\equiv 0$,

(a.1) take $N=S$, consider a total order $i_1,\ldots,i_s$, $s=|S|$ of elements of S and put $\mathsf{c}(\{i_1,i_s\})=1$, $\mathsf{c}(\{i_r,i_{r+1}\})=1$ for $r=1,\ldots s-1$, while c vanishes for remaining components and $\mathsf{a}\equiv 0$.

4 Necessity of the inequalities

In this section, we show that the inequalities from § 3.2 are necessarily valid for the vector codes from Definition 3.1. This allows us to interpret the inequalities in graphical terms.

4.1 Auxiliary observations

To verify the necessity of the extension inequalities we will need the next technical lemma.

Lemma 4.1 *Let H be a hybrid graph over N and $S \subseteq N$, $|S| \geq 3$.*

(a) *If H_S has a super-terminal component, then for at least $|S|-1$ subsets $T \subset S$ with $|T| = |S|-1$, the graph H_T has a super-terminal component.*

(b) *Provided H is 3-acyclic, if, for at least 3 subsets $T \subset S$ with $|T| = |S|-1$ the graph H_T has a super-terminal component, then H_S has a super-terminal component.*

Proof: (a) If H_S has a super-terminal component K, then for each $j \in S \setminus K$, the graph $H_{S\setminus\{j\}}$ has K as a super-terminal component. If $|K| \geq 2$, then, for every $i \in K$, $H_{S\setminus\{i\}}$ has $K \setminus \{i\}$ as a super-terminal component.
(b) Assume that $k, m, \ell \in S$ are such that, for each $i \in \{k, m, \ell\}$, $H_{S\setminus\{i\}}$ has a super-terminal component. Since $H_{\{k,m,\ell\}}$ has no directed cycle, it has a node with no arrow directed towards it. Thus, without loss of generality assume that k has the property that $\neg(m \to k \text{ in } H)$ and $\neg(\ell \to k \text{ in } H)$. Now,

let K be a super-terminal component in $H_{S\setminus\{k\}}$.

If $K \cap \{m, \ell\} \neq \emptyset$ then we can assume without loss of generality that $\ell \in K$. Thus, we distinguish the following three cases:

$\boxed{\text{A.}}$ $K \cap \{m, \ell\} = \emptyset$,

$\boxed{\text{B.}}$ $K = \{\ell\}$,

$\boxed{\text{C.}}$ $\ell \in K$ and $|K| \geq 2$.

In case $\boxed{\text{A.}}$ we have $K \subseteq S\setminus\{k, m, \ell\}$ and K is a super-terminal component in $H_{S\setminus\{k,m,\ell\}}$. Let L be a super-terminal component in $H_{S\setminus\{\ell\}}$. Then, because there exists $i \in K \subseteq S \setminus \{k, m, \ell\}$ with $m \to i$ in H, necessarily $m \notin L$. Also, $k \notin L$, as otherwise, for $m \notin L$, we observe $m \to k$ in H, which contradicts the choice of k. Hence, $L \subseteq S \setminus \{k, m, \ell\}$ and L is a super-terminal component in $H_{S\setminus\{k,m,\ell\}}$. By its uniqueness, we get $L = K$. This implies that K is a super-terminal component in H_S.

In case $\boxed{\text{B.}}$, that is $K = \{\ell\}$, let M be a super-terminal component in $H_{S\setminus\{m\}}$. Then, because for every $j \in S \setminus \{k, \ell\}$ one has $j \to \ell$ in H, necessarily $M \subseteq \{k, \ell\}$. Necessarily $\ell \in M$, for otherwise $M = \{k\}$ and $\ell \to k$ in H, which contradicts the choice of k. Then we have two options:

- $M = \{\ell\}$, in which case $k \to \ell$ in H and $K = M = \{\ell\}$ is a super-terminal component in H_S,
- or $M = \{k, \ell\}$, in which case $k — \ell$ in H. To show that $M = \{k, \ell\}$ is a super-terminal component in H_S, it remains to verify $m \to k$ in H. To this end, consider a super-terminal component L in $H_{S\setminus\{\ell\}}$. As $\forall\, j \in S \setminus \{k, m, \ell\}$ one has $j \to k$ in H, we have $L \subseteq \{k, m\}$. This implies that $[k, m]$ is an edge in H. Because H is 3-acyclic and $m \to \ell — k$ in H, it implies $m \to k$ in H, which was desired.

Finally, in case $\boxed{\text{C.}}$ the assumption $|K| \geq 2$ implies that the set $K \setminus \{\ell\}$ is a super-terminal component in $H_{S\setminus\{k,\ell\}}$. Now, let L be a super-terminal component in $H_{S\setminus\{\ell\}}$. We distinguish two subcases:

$\boxed{\text{C.1.}}$ If $k \notin L$, then L is a super-terminal component in $H_{S\setminus\{k,\ell\}}$, and, by its uniqueness, $L = K \setminus \{\ell\}$. To show that that K is a super-terminal component in H_S it remains to verify $k \to \ell$ in H. To this end, consider a super-terminal component M in $H_{S\setminus\{m\}}$. Because $\forall\, j \in S \setminus (\{k\} \cup K)$ one has $j \to \ell$ in H, we have $M \subseteq \{k\} \cup (K \setminus \{m\})$. As $K \setminus \{m\}$ is complete in H, there are three options for M, namely $M = \{k\}$, $M = K \setminus \{m\}$ and $M = \{k\} \cup (K \setminus \{m\})$. In each of these 3 cases, $[k, \ell]$ is an edge in H. There is at least one $i \in L \subseteq K$, and one has then $k \to i - \ell$ in H. Because H is 3-acyclic, it implies $k \to \ell$ in H, which was needed.

$\boxed{\text{C.2.}}$ If $k \in L$, then necessarily $m \in L$, for otherwise one has $m \to k$ in H, which contradicts the choice of k. Hence, $L \setminus \{k\}$ is a super-terminal component in $H_{S\setminus\{k,\ell\}}$, and, by its uniqueness, $K \setminus \{\ell\} = L \setminus \{k\} = K \cap L$. As $m \in K \cap L$, we have $k - m$ and $m - \ell$ in H. Let M be a super-terminal component in $H_{S\setminus\{m\}}$. Observe that $M \subseteq K \cup L$, because, $\forall\, j \in S \setminus (K \cup L)$, one has $j \to k$ in H. The next observation is that $k \in M$. Indeed, otherwise for some $i \in M \subseteq (K \cup L) \setminus \{m\}$ one has $k \to i$ in H, which means that $k \to i - m - k$ in H contradicting that H is 3-acyclic. Analogously, we derive $\ell \in M$ (exchange of k and ℓ). Because $k, \ell \in M$, we have $k - \ell$ in H, which implies that $K \cup L$ is complete in H_S. This allows to derive that $K \cup L$ is a super-terminal component in H_S.

Thus, in all cases (and subcases) we showed that H_S has a super-terminal component. □

4.2 Necessity proof

We show now that the inequalities from § 3.2 are valid for vector-codes introduced in § 3.1.

Lemma 4.2 *Let H be a hybrid graph over N which is 3-acyclic and has no flags.*

(i) *Then the inequalities (b.1)-(b.2), (c.1)-(c.5) and (e.1)-(e.2) are valid for the vector $(\mathsf{a}_H, \mathsf{c}_H)$ defined in (9).*

(ii) *If H is a chain graph without flags equivalent to an acyclic directed acyclic graph, then, moreover, the inequalities (a.1) hold.*

Proof: The inequalities (b.1)-(b.2) and (c.2) are evident from (9).

(c.1) If both terms on the left-hand side (LHS) of $\mathsf{a}_H(i \to j) + \mathsf{a}_H(j \to i) \leq \mathsf{c}_H(ij)$ vanish, then one has LHS $= 0 \leq$ RHS, where RHS means the right-hand side. Assume that at least one term on LHS is non-zero, say $\mathsf{a}_H(i \to j) = 1$. Then $i \to j$ in H, and because H is a hybrid graph, $\neg(j \to i$ in $H)$. Thus, the other term on LHS vanishes: $\mathsf{a}_H(j \to i) = 0$. However, $[i, j]$ is an edge in H then, which means, by (10), $\mathsf{c}_H(ij) = 1$. Hence, LHS $= 1 =$ RHS.

(c.3) If the LHS of $2 \cdot \mathsf{c}_H(ijk) \leq 2 \cdot \mathsf{c}_H(ij) + \mathsf{a}_H(i \to k) + \mathsf{a}_H(j \to k)$ vanishes, the inequality clearly holds. If $c_H(ijk) = 1 = \mathsf{c}_H(ij)$ then LHS $= 2 \leq$ RHS. Only in case $\mathsf{c}_H(ijk) = 1$ and $\mathsf{c}_H(ij) = 0$ we observe by (11) that $i \to k \leftarrow j$ is an immorality in H. Hence, LHS $= 2 =$ RHS.

(c.4) If at least one term on the LHS of $\mathsf{a}_H(i\to j)+\mathsf{c}_H(jk)\leq 1+\mathsf{c}_H(ijk)+\mathsf{a}_H(j\to k)$ vanishes LHS $\leq 1\leq$ RHS. Assume $\mathsf{a}_H(i\to j)=\mathsf{c}_H(jk)=1$, that is, $i\to j$ in H and $[j,k]$ is an edge in H. If $\mathsf{a}_H(j\to k)=1$ then LHS $=2\leq$ RHS. If $\mathsf{a}_H(j\to k)=0$, then, since H is a hybrid graph, either $j\leftarrow k$ in H or j — k in H. Because H has no flag (of the form $i\to j$ — k), in both these sub-cases one has $\mathsf{c}_H(ijk)=1$ by (11). Thus, LHS $=2=$ RHS.

(c.5) If at least one term on the LHS of $\mathsf{a}_H(i\to j)+\mathsf{c}_H(jk)+\mathsf{c}_H(ik)\leq 2+\mathsf{a}_H(i\to k)+\mathsf{a}_H(k\to j)$ vanishes then LHS $\leq 2\leq$ RHS. If all three of them equal 1 then $i\to j$ in H and $[j,k]$,$[k,i]$ are both edges in H. As H is 3-acyclic one necessarily has either $i\to k$ in H or $k\to j$ in H. In particular, LHS $=3\leq$ RHS.

(e.1) If at most two terms on the LHS of $\sum_{i\in S}\mathsf{c}_H(S\setminus\{i\})\leq 2+(|S|-2)\cdot\mathsf{c}_H(S)$ are non-zero then LHS $\leq 2\leq$ RHS. If at least three of them are non-zero then, by (9), for at least 3 subsets $T\subset S$ with $|T|=|S|-1$, the graph H_T has a super-terminal component. By Lemma 4.1(b), H_S has a super-terminal component, and, therefore $\mathsf{c}_H(S)=1$. Hence, LHS $\leq |S|=$ RHS.

(e.2) If the LHS of $(|S|-1)\cdot\mathsf{c}_H(S)\leq\sum_{i\in S}\mathsf{c}_H(S\setminus\{i\})$ vanishes then LHS $=0\leq$ RHS. However, if $\mathsf{c}_H(S)=1$ then, by (9), H_S has a super-terminal component. By Lemma 4.1(a) observe that there exists at least $|S|-1$ subsets $T\subset S$ with $|T|=|S|-1$ with $\mathsf{c}_H(T)=1$. Thus, by (9), LHS $=|S|-1\leq$ RHS.

Thus, the part (i) of Lemma 4.2 was proved. To verify the part (ii) assume that H is a chain graph without flags equivalent to an acyclic directed acyclic graph G. If follows from (9) and Lemma 2.2 that c_H equals to the characteristic imset c_G for G. Thus, we need to verify the inequality

(a.1) $\sum_{T\subseteq S,\,|T|\geq 2}\mathsf{c}_G(T)\cdot(-1)^{|T|}\leq |S|-1$ for any acyclic directed graph G over N.

Let us fix $S\subseteq N$, $|S|\geq 4$. As G_S is acyclic, there exists a total order ρ of all nodes in S consistent with the direction of arrows in G_S. For every $j\in S$ let us put

$$\mathcal{T}(j)=\{T\subseteq\{j\}\cup B^{\rho}_j;\ |T|\geq 2,\ j\in T\}\quad\text{where } B^{\rho}_j \text{ is the set of predecessors of } j \text{ in } \rho.$$

Clearly, the classes $\mathcal{T}(j)$, $j\in S$ define a partitioning of the class $\{T\subseteq S;\ |T|\geq 2\}$; one has $\mathcal{T}(i)=\emptyset$ for the first node i in ρ (= the one with $B^{\rho}_i=\emptyset$). The point is that, for every $j\in S\setminus\{i\}$, one has

$$\sum_{T\in\mathcal{T}(j)}\mathsf{c}_G(T)\cdot(-1)^{|T|}\leq 1\,. \tag{12}$$

Indeed, by (1), observe that, for $T\in\mathcal{T}(j)$, one has $\mathsf{c}_G(T)=1$ iff $T=\{j\}\cup K$, where $K\subseteq pa_G(j)\cap S$. In particular, the sum on the LHS of (12) is

$$\sum_{\emptyset\neq K\subseteq pa_G(j)\cap S}(-1)^{|K|+1}=\begin{cases}0 & \text{if } pa_G(j)\cap S=\emptyset,\\ 1 & \text{if } pa_G(j)\cap S\neq\emptyset.\end{cases}$$

Clearly, (a.1) can be obtained by summing (12) for $j\in S\setminus\{i\}$. □

Note that (a.1) is, in fact, a transformed cluster inequality (3): the following equality can be derived using (4) with some effort:

$$|S|-\sum_{T\subseteq S,\,|T|\geq 2}\mathsf{c}_G(T)\cdot(-1)^{|T|}=\sum_{i\in S}\ \sum_{B\subseteq N\setminus S}\eta_G(i|B)\,;$$

for details see the proof of Lemma 16 in [21].

Now, it is clear from the above proof of Lemma 4.2 what is the graphical interpretation of the inequalities from § 3.2:

(c.1) $\mathsf{a}_H(i \to j) + \mathsf{a}_H(j \to i) \leq \mathsf{c}_H(ij)$ means that if $i \to j$ or $j \to i$ (are encoded in a_H as arrows) then $[i,j]$ is (encoded in c_H as) an edge,

(c.2) $\mathsf{c}_H(ij) \leq 1$ means, together with (c.1), that there is no bi-directed edge between i and j (= one cannot have simultaneously $i \to j$ in H and $j \to i$ in H),

(c.3) $2 \cdot \mathsf{c}_H(ijk) \leq 2 \cdot \mathsf{c}_H(ij) + \mathsf{a}_H(i \to k) + \mathsf{a}_H(j \to k)$ allows to conclude that if $\mathsf{c}_H(ijk) = 1$ then H_{ijk} has a super-terminal component,

(c.4) $\mathsf{a}_H(i \to j) + \mathsf{c}_H(jk) \leq 1 + \mathsf{c}_H(ijk) + \mathsf{a}_H(j \to k)$ prevents H to have a flag $i \to j — k$,

(c.5) $\mathsf{a}_H(i \to j) + \mathsf{c}_H(jk) + \mathsf{c}_H(ik) \leq 2 + \mathsf{a}_H(i \to k) + \mathsf{a}_H(k \to j)$ says that H has not a semi-directed 3-cycle of the form i, j, k, i with $i \to j$,

(e.1) $\sum_{i \in S} \mathsf{c}_H(S \setminus \{i\}) \leq 2 + (|S| - 2) \cdot \mathsf{c}_H(S)$ encodes the implication in Lemma 4.1(b),

(e.2) $(|S| - 1) \cdot \mathsf{c}_H(S) \leq \sum_{i \in S} \mathsf{c}_H(S \setminus \{i\})$ encodes the implication in Lemma 4.1(a),

(a.1) $\sum_{T \subseteq S, |T| \geq 2} \mathsf{c}_H(T) \cdot (-1)^{|T|} \leq |S| - 1$ means, loosely said, that the graph H_S has at least one initial node. The point is that all inequalities (a.1) together ensure the existence of an equivalent acyclic directed graph G, as shown in the following section.

5 Sufficiency of the inequalities

In this section, we prove that the inequalities from § 3.2 provide an LP relaxation for (the convex hull of) the set of codes of chain graphs without flags equivalent to acyclic directed graphs.

5.1 Auxiliary lemmas

We start with a couple of auxiliary observations.

Lemma 5.1 *Let* (a, c) *be a vector with integer components satisfying the inequalities (b.1)-(b.2), (c.1)-(c.5) and (e.1)-(e.2). Then (e.2) also holds for* $|S| = 3$, *that is,*

$$2 \cdot c(ijk) \leq \mathsf{c}(ij) + \mathsf{c}(ik) + \mathsf{c}(jk) \,, \tag{13}$$

and (a, c) *is a zero-one vector.*

Proof: To show (13) let us fix the set $S = ijk$ and sum up all three corresponding inequalities (c.3) and then use three-times (c.1):

$$6 \cdot c(ijk) \leq 2 \cdot c(ij) + 2 \cdot c(ik) + 2 \cdot c(jk) + \mathsf{a}(i \to j) + \mathsf{a}(j \to i)$$
$$+ \mathsf{a}(i \to k) + \mathsf{a}(k \to i) + \mathsf{a}(j \to k) + \mathsf{a}(k \to j) \leq 3 \cdot c(ij) + 3 \cdot c(ik) + 3 \cdot c(jk) \,.$$

Divide it by 3 and get (13). The next observation is that (a, c) is non-negative. The inequality $0 \leq \mathsf{c}(S)$ for $|S| = 2$ follows from (b.1) and (c.1). The inequality $0 \leq \mathsf{c}(S)$ for

$|S| \geq 5$ can be derived by induction on $|S|$ from (b.2) and (e.1): if $|S| \geq 5$ then by the induction premise and (e.1) one has

$$0 \leq 2 + (|S|-2) \cdot \mathsf{c}(S), \quad \text{which implies} \quad -1 < \frac{-2}{|S|-2} \leq \mathsf{c}(S)\,.$$

However, because $\mathsf{c}(S) \in \mathbb{Z}$, it implies $0 \leq \mathsf{c}(S)$.

Finally, we show that the components of (a, c) are at most 1. As concerns the a-part, by (b.1), (c.1) and (c.2) one has $\mathsf{a}(i \to j) \leq \mathsf{a}(i \to j) + \mathsf{a}(j \to i) \leq c(ij) \leq 1$. As concerns the c-part, we prove $\mathsf{c}(S) \leq 1$ by induction on $|S|$ from (c.2) using (e.2) for $|S| \geq 3$, which involves (13). Specifically, consider $|S| > 2$ and write by (e.2) and the induction premise

$$(|S|-1) \cdot \mathsf{c}(S) \leq \sum_{i \in S} \mathsf{c}(S \setminus \{i\}) \leq \sum_{i \in S} 1 = |S|\,. \quad \text{Hence,} \quad \mathsf{c}(S) \leq \frac{|S|}{|S|-1} = 1 + \frac{1}{|S|-1} < 2\,.$$

However, because $\mathsf{c}(S) \in \mathbb{Z}$, it implies $\mathsf{c}(S) \leq 1$. □

Lemma 5.2 *Let* c,c' *be zero-one vectors satisfying (e.1)-(e.2). Provided* $\mathsf{c}(S) = \mathsf{c}'(S)$ *for* $S \subseteq N$ *with* $2 \leq |S| \leq 3$ *one has* $\mathsf{c} = \mathsf{c}'$.

Proof: We prove $\mathsf{c}(S) = \mathsf{c}'(S)$ by induction on $|S|$. Since they are both zero-one vectors and exchangeable, it it enough to show, for $|S| \geq 4$, the implication $\mathsf{c}(S) = 1 \;\Rightarrow\; \mathsf{c}'(S) = 1$. First, by (e.2) applied to c, we observe $3 \leq |S| - 1 = (|S|-1) \cdot \mathsf{c}(S) \leq \sum_{i \in S} \mathsf{c}(S \setminus \{i\})$. By the induction premise the same holds for c' and one can write using (e.1) applied to c':

$$3 \leq \sum_{i \in S} \mathsf{c}'(S \setminus \{i\}) \leq 2 + (|S|-2) \cdot \mathsf{c}'(S)\,. \quad \text{That means} \;\; 0 < \frac{1}{|S|-2} \leq \mathsf{c}'(S)\,,$$

which, because $c'(S) \in \{0,1\}$, implies $\mathsf{c}'(S) = 1$. □

5.2 Sufficiency proof

Here is the desired result.

Lemma 5.3 *Given a finite non-empty set* N*, let* (a, c) *be a vector with integer components satisfying the inequalities (b.1)-(b.2), (c.1)-(c.5) and (e.1)-(e.2).*

(i) *Then there exists a (unique) hybrid graph* H *over* N *which is 3-acyclic and has no flags such that* $\mathsf{a} = \mathsf{a}_H$ *and* $\mathsf{c} = \mathsf{c}_H$.

(ii) *If, moreover, the inequalities (a.1) hold for* c *then* H *is a chain graph without flags equivalent to an acyclic directed graph over* N.

Proof: By Lemma 5.1, (a, c) is a zero-one vector. Let us define a hybrid graph H over N:

$$\begin{aligned} i \to j \;\; \text{in } H \quad &\Leftrightarrow \quad \mathsf{a}(i \to j) = 1, \\ i - j \;\; \text{in } H \quad &\Leftrightarrow \quad \mathsf{c}(ij) = 1 \;\&\; \mathsf{a}(i \to j) = 0 \;\&\; \mathsf{a}(j \to j) = 0\,. \end{aligned} \qquad (14)$$

The definition is correct, since, by (c.1) and (c.2), $\mathsf{a}(i \to j) + \mathsf{a}(j \to i) \leq \mathsf{c}(ij) \leq 1$. Note

$$\mathsf{c}(ij) = 1 \;\Leftrightarrow\; [i,j] \text{ is an edge in } H \;\Leftrightarrow\; H_{ij} \text{ has a super-terminal component.} \tag{15}$$

Observe that H is 3-acyclic. Indeed, assume for a contradiction that H has a semi-directed cycle $\rho : i, j, k, i$ with $i \to j$. Hence, by (14), (15) and (c.5) we have

$$3 = \mathsf{a}(i \to j) + \mathsf{c}(jk) + \mathsf{c}(ik) \leq 2 + \mathsf{a}(i \to k) + \mathsf{a}(k \to j),$$

that means $1 \leq \mathsf{a}(i \to k) + \mathsf{a}(k \to j)$, which implies, by (14), that either $i \to k$ in H or $k \to j$ in H. This contradict the assumption ρ is a semi-directed cycle.

The next step is to show that, for distinct $i, j, k \in N$,

$$\mathsf{c}(ijk) = 1 \;\Leftrightarrow\; H_{ijk} \text{ has a super-terminal component.} \tag{16}$$

For this purpose, realize that, because H is 3-acyclic, the graph H_{ijk} has a super-terminal component iff it is either quasi-complete or an immorality. For $\Rightarrow$ direction, if $\mathsf{c}(ijk) = 1$ and H_{ijk} is not quasi-complete, then at least one edge in H_{ijk} is missing. Assume without loss of generality that it is $[i,j]$, which means, by (15), $\mathsf{c}(ij) = 0$. Write by (c.3)

$$2 = 2 \cdot \mathsf{c}(ijk) \leq 2 \cdot \mathsf{c}(ij) + \mathsf{a}(i \to k) + \mathsf{a}(j \to k) = \mathsf{a}(i \to k) + \mathsf{a}(j \to k)\,.$$

This implies that both $\mathsf{a}(i \to k) = 1$ and $\mathsf{a}(j \to k) = 1$; by (14), $i \to k \leftarrow j$ is an immorality. For $\Leftarrow$ direction, in case H_{ijk} is quasi-complete, write by (15) and (e.1) for $S = ijk$:

$$3 = \mathsf{c}(ij) + \mathsf{c}(ik) + \mathsf{c}(jk) \leq 2 + \mathsf{c}(ijk),$$

which implies the desired conclusion $\mathsf{c}(ijk) = 1$. If H_{ijk} is an immorality, say $i \to j \leftarrow k$, then realize that $\neg(j \to k \text{ in } H)$ and observe by (14) that $\mathsf{a}(i \to j) = 1$ and $\mathsf{a}(j \to k) = 0$; while $\mathsf{c}(jk) = 1$ by (15). This allows us to write by (c.4)

$$2 = \mathsf{a}(i \to j) + \mathsf{c}(jk) \leq 1 + \mathsf{c}(ijk) + \mathsf{a}(j \to k) = 1 + \mathsf{c}(ijk),$$

implying $\mathsf{c}(ijk) = 1$. Thus, (16) has been verified.

The next observation is that H has no flag. Assume for a contradiction it has a flag $i \to j - k$. As $\neg(j \to k \text{ in } H)$, by (14), $\mathsf{a}(i \to j) = 1$ and $\mathsf{a}(j \to k) = 0$. By (15) and (c.4)

$$2 = \mathsf{a}(i \to j) + \mathsf{c}(jk) \leq 1 + \mathsf{c}(ijk) + \mathsf{a}(j \to k) = 1 + \mathsf{c}(ijk).$$

Hence, $\mathsf{c}(ijk) = 1$, which, by (16), contradicts the assumption $i \to j - k$ is a flag in H.

Thus, consider the 3-acyclic hybrid graph H without flags. By Lemma 4.2(i) the vector c_H given by (9) is a zero-one vector satisfying the conditions (e.1)-(e.2). By (15) and (16), we have $\mathsf{c}(S) = \mathsf{c}_H(S)$ for any $S \subseteq N$, $2 \leq |S| \leq 3$. As c also satisfies (e.1)-(e.2), by Lemma 5.2, we have $\mathsf{c} = \mathsf{c}_H$. By (9) and (14), we also have $\mathsf{a} = \mathsf{a}_H$, which concludes the proof of the (i)-part of Lemma 5.3.

As concerns the (ii)-part, assume c satisfies (a.1). As mentioned in § 2.1, to show that H is a chain graph, it is enough to verify it has no semi-directed chordless cycle. Since H is 3-acyclic, we already know it has no semi-directed cycle of the length 3. Thus, assume for a contradiction it has a semi-directed chordless cycle $\rho : i_0, i_1, \ldots, i_m = i_0$ of the length $m \geq 4$ with $i_0 \to i_1$ in H. Because H has no flag, we observe that ρ has to be a directed cycle, that is, $i_r \to i_{r+1}$ in H for $r = 0, 1, \ldots, m-1$. Put $S = \{i_1, \ldots, i_m\}$. By the (i)-part, $\mathsf{c} = \mathsf{c}_H$; thus, using (9) observe that the only subsets $T \subseteq S$, $|T| \geq 2$ with $\mathsf{c}(T) = 1$ are the edges of the cycle ρ. In particular, if we substitute into (a.1) we get

$$|S| = \sum_{T \subseteq S,\, |T| \geq 2} \mathsf{c}(T) \cdot (-1)^{|T|} \leq |S| - 1\,,$$

which is a contradiction. Thus, H cannot have a semi-directed cycle of the length $m \geq 4$.

As mentioned in § 2.3, to show that H is equivalent to an acyclic directed graph, it is enough to show it has no undirected chordless cycle of the length $m \geq 4$. The proof by a contradiction is analogous to the case of a semi-directed cycle. If $\rho : i_0 - i_1 - \ldots - i_m = i_0$ is such a cycle, put $S = \{i_1, \ldots, i_m\}$ and get a contradiction with (a.1) in the same way. □

Thus, by combining Lemmas 4.2 and 5.3 we get.

Theorem 5.1 *Given a finite non-empty set N, a vector (a, c) with integer components satisfies the inequalities (b.1)-(b.2), (c.1)-(c.5), (e.1)-(e.2) and (a.1) if and only if there exists (uniquely determined) chain graph H over N without flags equivalent to an acyclic directed graph such that $(\mathsf{a}, \mathsf{c}) = (\mathsf{a}_H, \mathsf{c}_H)$.*

5.3 Modification of acyclicity inequalities

In this section, we show that the acyclicity inequalities (a.1) can be modified. Specifically, they can be replaced by a simplified version

(a.1*) $\forall\, S \subseteq N,\ |S| \geq 4 \qquad \sum_{T \subseteq S,\, |T|=2} \mathsf{c}(T) - \sum_{T \subseteq S,\, |T|=3} \mathsf{c}(T) \leq |S| - 1,$

which may perhaps appear to be easier to implement. However, this change formally leads to a different LP relaxation of the same polytope (= of the convex hull of the considered set of codes). Note that Lindner [12] named (a.1*) the *DAG inequalities* and included them in the LP relaxation of her polytope, which was different (from the polytope considered here).

Theorem 5.2 *Given a finite non-empty set N, a vector (a, c) with integer components satisfies the inequalities (b.1)-(b.2), (c.1)-(c.5), (e.1)-(e.2) and (a.1*) if and only if there exists a chain graph H over N without flags equivalent to an acyclic directed graph such that $(\mathsf{a}, \mathsf{c}) = (\mathsf{a}_H, \mathsf{c}_H)$.*

Proof: As concerns the necessity proof, by Lemma 4.2(i), it is enough to show that (a.1*) holds for such $(\mathsf{a}_H, \mathsf{c}_H)$. The procedure is analogous to the case (a.1) in the proof of Lemma 4.2(ii). Basically, we need to verify that, for any acyclic directed graph G over N,

$$\text{(a.1*)} \quad \sum_{T \subseteq S,\, |T|=2} \mathsf{c}_G(T) - \sum_{T \subseteq S,\, |T|=3} \mathsf{c}_G(T) \leq |S| - 1 \qquad \text{whenever } S \subseteq N,\ |S| \geq 4.$$

Again, we fix S and a total order ρ of nodes in S consistent with the direction of arrows in G_S. Let i denote the first node in ρ and consider the same partitioning of the class $\{T \subseteq S;\ |T| \geq 2\}$ into subclasses $\mathcal{T}(j)$, $j \in S \setminus \{i\}$. In this case, however, the following collection of inequalities is derived

$$\sum_{T \in \mathcal{T}(j),\, |T|=2} \mathsf{c}_G(T) - \sum_{T \in \mathcal{T}(j),\, |T|=3} \leq 1\,. \qquad (17)$$

Indeed, for any $T \in \mathcal{T}(j)$, one has $\mathsf{c}_G(T) = 1$ iff $T = \{j\} \cup K$, where $K \subseteq pa_G \cap S$. Therefore, the sum on the LHS of (17) is as follows:

$$\sum_{K \subseteq pa_G(j) \cap S,\, |K|=1} 1 - \sum_{K \subseteq pa_G(j) \cap S,\, |K|=2} 1 = \begin{cases} 0 & \text{if } pa_G(j) \cap S = \emptyset, \\ 1 & \text{if } 1 \leq |pa_G(j) \cap S| \leq 2, \\ -\frac{\ell \cdot (\ell-3)}{2} \leq 0 & \text{if } 3 \leq \ell \equiv |pa_G(j) \cap S|. \end{cases}$$

Of course, (a.1*) can be obtained by summing (17) for $j \in S \setminus \{i\}$.

For the sufficiency use Lemma 5.3(i) and then modify the proof of Lemma 5.3(ii): the same arguments hold if (a.1) is replaced by (a.1*) there. □

An alternative proof of the necessity of (a.1*) is to derive it as a consequence of (a.1). One can show that the following inequality

$$\sum_{T \subseteq S,\, |T| \geq 4} \mathsf{c}_H(T) \cdot (-1)^{|T|} \geq 0 \qquad \text{where } S \subseteq N,\ |S| \geq 4 \tag{18}$$

is valid for any a chain graph H without flags equivalent to an acyclic directed acyclic graph; by subtracting this inequality from (a.1) one gets (a.1*).

A natural question is what is the relation of the LP relaxations (of the same polytope) from Theorems 5.1 and 5.2. In case $|N| = 4$, (18), namely $\mathsf{c}(N) \geq 0$, follows from (b.2). Therefore, in case $|N| = 4$, the LP relaxation with (a.1) defines a smaller polyhedron than the one with (a.1*). The following example shows that the inclusion is strict.

Example 5.1 Consider $N = \{a, b, c, d\}$, an undirected graph H over N formed by a chordless cycle $a — b — c — d — a$, and an acyclic directed graph G over N with (three) arrows, directed from nodes a, b, c into d. Note that, by Lemma 4.2(i), both c_H and c_G satisfy (b.1)-(b.2), (c.1)-(c.5) and (e.1)-(e.2). In particular, their convex combination $\mathsf{c} := \frac{3}{4} \cdot \mathsf{c}_H + \frac{1}{4} \cdot \mathsf{c}_G$ satisfies those inequalities, too. Moreover, c also satisfies (a.1*) for $S = N$ with equality:

$$\sum_{|T|=2} \mathsf{c}(T) - \sum_{|T|=3} \mathsf{c}(T) = 2 \cdot 1 + 2 \cdot \frac{3}{4} + \frac{1}{4} - 3 \cdot \frac{1}{4} = 3 \leq |N| - 1\,.$$

However, in (a.1) an additional term $\mathsf{c}(N) = \frac{1}{4}$ is on the LHS of the inequality, for which reason it does not hold for c.

It looks like one cannot expect that (a.1*) is implied by the inequalities from Theorem 5.1 in general, for $|N| \geq 5$. Nevertheless, the LP relaxation based on (a.1) always gives, in a certain sense, a tighter approximation of the considered polytope. More specifically, the face defined by (a.1) contains more lattice points than the corresponding face defined by (a.1*). This can be derived from the following auxiliary observation.

Lemma 5.4 *Let N be a finite non-empty set.*

(i) *Consider the polyhedron specified by the inequalities (b.1)-(b.2), (c.1)-(c.5), (e.1)-(e.2) and (a.1). Then a vector with integer components (a, c) satisfies (a.1) for $S \subseteq N$, $|S| \geq 4$ with the equality iff $(\mathsf{a}, \mathsf{c}) = (\mathsf{a}_H, \mathsf{c}_H)$, where H is equivalent to an acyclic directed graph G such that there is only one initial node in G_S (= just one $i \in S$ with $pa_G(i) \cap S = \emptyset$).*

(ii) *Consider the polyhedron specified by the inequalities (b.1)-(b.2), (c.1)-(c.5), (e.1)-(e.2) and (a.1*). Then a vector with integer components* (a,c) *satisfies (a.1*) for* $S \subseteq N$, $|S| \geq 4$ *with the equality iff* $(\mathsf{a},\mathsf{c}) = (\mathsf{a}_H,\mathsf{c}_H)$*, where* H *is equivalent to an acyclic directed graph* G *such that* G_S *has the property that, for every* $j \in S$ *but one,* $1 \leq |pa_G(j) \cap S| \leq 2$.

Proof: As concerns **(i)**, by Theorem 5.1, every lattice point in that polyhedron is a code $(\mathsf{a}_H,\mathsf{c}_H)$ of the respective graph H over N. Now, repeat the consideration in the proof of Lemma 4.2(ii). Specifically, it is shown there that (a.1) is obtained by summing inequalities (12). In particular, (a.1) holds with the equality iff the equality holds in (12) for every $j \in S\setminus\{i\}$. This happens (see the argument in the proof of Lemma 4.2(ii)) if $pa_G(j)\cap S \neq \emptyset$ for any such j. That means that i is the only initial node in G_S.

As concerns **(ii)**, the procedure is analogous. We repeat the consideration from the proof of Theorem 5.2, where it is shown that (a.1*) is obtained by summing inequalities (17). In particular, (a.1*) holds with the equality iff the equality holds in (17) for every $j \in S \setminus \{i\}$, that is, when $1 \leq |pa_G(j) \cap S| \leq 2$ for any such j. □

Thus, from Lemma 5.4, one can easily derive the following.

Corollary 5.1 *Let* N *be a finite non-empty set and* $S \subseteq N$, $|S| \geq 4$. *Then, every lattice point within the polyhedron specified by (b.1)-(b.2),(c.1)-(c.5), (e.1)-(e.2) and (a.1*) which belongs the face (of that polyhedron) defined by (a.1*) with* S *also belongs to the face defined by (a.1) for* S *of the polyhedron specified by (b.1)-(b.2),(c.1)-(c.5),(e.1)-(e.2) and (a.1).*

The fact that the face defined by (a.1) for S contains many more other lattice points may play an important role in solving the respective ILP problems. One can perhaps expect less frequent need for the application of branch-and-bound procedure. Actually, the results from [21] suggest that (a.1) should be facet-defining inequalities for the considered polytope (= the convex hull of the set of codes).

6 First phase: getting the characteristic imset

In this section we formulate the first ILP problem in its pure form. Its solution should be the respective characteristic imset.

6.1 Score as an affine function

It was observed in Lemma 1 of [9], provided $\mathcal{Q}$ is additively decomposable and score equivalent, it can be viewed as (the restriction of) an affine function of the characteristic imset c_G. Specifically, we have

$$\mathcal{Q}(G,D) \;=\; \mathcal{Q}(G^{\emptyset},D) + \sum_{S\subseteq N,\,|S|\geq 2} r_D^{\mathcal{Q}}(S)\cdot \mathsf{c}_G(S)\,, \tag{19}$$

where $G^{\emptyset}$ is the empty graph over N (= the graph without edges) and $r_D^{\mathcal{Q}}$ is the so-called *revised data vector* (relative to $\mathcal{Q}$). It is uniquely determined (by $\mathcal{Q}$ and D) and can be obtained from the local scores $q_D(*|*)$ as described below.

Lemma 6.1 *Let $\mathcal{Q}$ be a score equivalent and additively decomposable criterion with local scores $q_D(*|*)$. First, the local scores are standardized:*

$$\hat{q}_D(i|B) = q_D(i|B) - q_D(i|\emptyset) \qquad \text{for } i \in N,\ B \subseteq N \setminus \{i\}. \tag{20}$$

For every $T \subseteq N$, $|T| \geq 2$, consider a total order ρ of elements in T; denote, for $i \in T$, by the symbol B_i^ρ the set of predecessors of $i \in T$ in ρ. Second, we observe that the expression

$$t_D^{\mathcal{Q}}(T) = \sum_{i \in T} \hat{q}_D(i \,|\, B_i^\rho) \qquad \text{for } T \subseteq N,\ |T| \geq 2\,, \tag{21}$$

does not depend on the choice of ρ. Third, this makes correct the following definition:

$$r_D^{\mathcal{Q}}(S) = \sum_{T \subseteq S,\, |T| \geq 2} (-1)^{|S \setminus T|} \cdot t_D^{\mathcal{Q}}(T) \qquad \text{for } S \subseteq N,\ |S| \geq 2\,. \tag{22}$$

Then, the formula (19) holds.

The vector $t_D^{\mathcal{Q}}$ from (21) have been introduced earlier in [18] and called the *data vector* relative to $\mathcal{Q}$; note that, if one puts $t_D^{\mathcal{Q}}(T) = 0$ for $T \subseteq N$, $|T| \leq 1$, then (21) implies that,

$$\hat{q}_D(i|B) = t_D^{\mathcal{Q}}(\{i\} \cup B) - t_D^{\mathcal{Q}}(B) \qquad \text{for any } i \in N \text{ and } B \subseteq N \setminus \{i\}\,. \tag{23}$$

Proof: It follows from (5) and (20) that, for any G and D, one has

$$\mathcal{Q}(G, D) = \sum_{i \in N} q_D(i \,|\, \emptyset) + \sum_{i \in N} \{\, q_D(i \,|\, \mathrm{pa}_G(i)) - q_D(i \,|\, \emptyset)\} = \mathcal{Q}(G^\emptyset, D) + \sum_{i \in N} \hat{q}_D(i \,|\, \mathrm{pa}_G(i))\,.$$

Given $T \subseteq N$, $|T| \geq 2$ and a total order ρ of its elements, let G^ρ be the acyclic directed graph over N in which $j \to i$ iff $i, j \in T$ and $j \in B_i^\rho$. Thus, we have

$$\sum_{i \in T} \hat{q}_D(i \,|\, B_i^\rho) = \sum_{i \in N} \hat{q}_D(i \,|\, \mathrm{pa}_{G^\rho}(i)) = \mathcal{Q}(G^\rho, D) - \mathcal{Q}(G^\emptyset, D),$$

which implies, by the assumption that $\mathcal{Q}$ is score equivalent, that the right-hand side of (21) does not depend on the choice of ρ. Because of the first formula in this proof, to prove (19), it is enough to verify

$$\sum_{S \subseteq N,\, |S| \geq 2} r_D^{\mathcal{Q}}(S) \cdot \mathsf{c}_G(S) = \sum_{i \in N} \sum_{B \subseteq N \setminus \{i\}} \eta_G(i|B) \cdot \hat{q}_D(i|B) \equiv \sum_{i \in N} \hat{q}_D(i \,|\, \mathrm{pa}_G(i))\,.$$

To show it, we use the symbol $\delta(\star\star)$ for the zero-one indicator of the validity of a statement $\star\star$, substitute (4) into the left-hand side expression and change the order of summation:

$$\begin{aligned} \sum_{S \subseteq N,\, |S| \geq 2} r_D^{\mathcal{Q}}(S) \cdot \mathsf{c}_G(S) &= \sum_{S \subseteq N,\, |S| \geq 2} r_D^{\mathcal{Q}}(S) \cdot \sum_{i \in N} \sum_{B \subseteq N \setminus \{i\}} \delta(i \in S) \cdot \delta(S \setminus \{i\} \subseteq B) \cdot \eta_G(i|B) \\ &= \sum_{i \in N} \sum_{B \subseteq N \setminus \{i\}} \eta_G(i|B) \cdot \underbrace{\sum_{S \subseteq N,\, |S| \geq 2} \delta(i \in S) \cdot \delta(S \setminus \{i\} \subseteq B) \cdot r_D^{\mathcal{Q}}(S)}_{\hat{q}_D(i|B)}\,. \end{aligned}$$

Thus, it remains to verify, for any fixed $i \in N$ and $B \subseteq N \setminus \{i\}$, that the indicated inner expression is indeed $\hat{q}_D(i|B)$. To this end, we substitute (22) into it, change the order of summation and re-write that a little bit to get what is desired:

$$\begin{aligned}
&\sum_{S\subseteq N,\,|S|\geq 2} \delta(i\in S)\cdot\delta(S\setminus\{i\}\subseteq B)\cdot r^{\mathcal{Q}}_D(S) = \\
&= \sum_{\substack{S\subseteq\{i\}\cup B\\ i\in S,\,|S|\geq 2}} r^{\mathcal{Q}}_D(S) = \sum_{\substack{S\subseteq\{i\}\cup B\\ i\in S,\,|S|\geq 2}} \;\sum_{\substack{T\subseteq S\\ |T|\geq 2}} (-1)^{|S\setminus T|}\cdot t^{\mathcal{Q}}_D(T) \\
&= \sum_{\substack{T\subseteq\{i\}\cup B\\ |T|\geq 2}} t^{\mathcal{Q}}_D(T)\cdot \sum_{\substack{S,\,T\subseteq S\subseteq\{i\}\cup B\\ i\in S}} (-1)^{|S\setminus T|} \\
&= \sum_{\substack{T\subseteq\{i\}\cup B\\ |T|\geq 2}} t^{\mathcal{Q}}_D(T)\cdot \sum_{\substack{S,\,T\subseteq S\subseteq\{i\}\cup B\\ i\in S}} (-1)^{\delta(i\notin T)}\cdot(-1)^{|S\setminus(\{i\}\cup T)|} \\
&= \sum_{T\subseteq\{i\}\cup B,\,|T|\geq 2} t^{\mathcal{Q}}_D(T)\cdot(-1)^{\delta(i\notin T)}\cdot \underbrace{\sum_{S,\,\{i\}\cup T\subseteq S\subseteq\{i\}\cup B} (-1)^{|S\setminus(\{i\}\cup T)|}}_{\delta(\,\{i\}\cup T=\{i\}\cup B\,)} \\
&= \delta(|B|\geq 1)\cdot t^{\mathcal{Q}}_D(\{i\}\cup B) - \delta(|B|\geq 2)\cdot t^{\mathcal{Q}}_D(B) \;=\; \hat{q}_D(i|B)\,.
\end{aligned}$$

To see the last equality use (23). □

Lemma 6.1 allows us to derive a direct formula for the components of the revised data vector on the basis of local scores.

Corollary 6.1 *Let $\mathcal{Q}$ be a score equivalent and additively decomposable criterion with local scores $q_D(*|*)$. Then, for any $S \subseteq N$, $|S| \geq 2$, one has*

$$r^{\mathcal{Q}}_D(S) = \sum_{K\subseteq R} (-1)^{|R\setminus K|}\cdot q_D(j|K) \qquad \textit{where } j\in S \textit{ and } R = S\setminus\{j\}\,. \tag{24}$$

In particular, the right-hand side of (24) does not depend on the choice of $j \in S$.

Proof: First, note that it is enough to prove (24) with standardized local scores $\hat{q}_D(*|*)$, given by (20); this is because $\sum_{K\subseteq R}(-1)^{|R\setminus K|}\cdot q_D(j|\emptyset) = 0$. That means, as $\hat{q}_D(j|\emptyset) = 0$, we are going to verify

$$r^{\mathcal{Q}}_D(S) = \sum_{\emptyset\neq K\subseteq R} (-1)^{|R\setminus K|}\cdot \hat{q}_D(j|K) \qquad \text{for any } S\subseteq N,\ |S|\geq 2,\ j\in S \text{ and } R = S\setminus\{j\}\,.$$

For this purpose choose a total ordering π of elements in S which ends by j, that is, $R = S \setminus \{j\}$ is the set of predecessors of j in π. For any $T \subseteq S$, $|T| \geq 2$ consider a total order ρ of elements in T induced by π. Then the set of predecessors of $i \in T$ is $\pi(i) \cap T$, where $\pi(i)$ denotes the set of predecessors of i in π. Now, we can substitute (21) into (22) and change the order of summation:

$$\begin{aligned} r_D^{\mathcal{Q}}(S) &= \sum_{T\subseteq S,\,|T|\geq 2} (-1)^{|S\setminus T|}\cdot t_D^{\mathcal{Q}}(T) = \sum_{T\subseteq S,\,|T|\geq 2} (-1)^{|S\setminus T|}\cdot \sum_{i\in T} \hat{q}_D(i\,|\,\pi(i)\cap T) \\ &= \sum_{i\in S} \sum_{T\subseteq S,\,|T|\geq 2,\,i\in T} (-1)^{|S\setminus T|}\cdot \hat{q}_D(i\,|\,\pi(i)\cap T) \\ &= \sum_{i\in S} \sum_{K\subseteq \pi(i)} \hat{q}_D(i|K)\cdot \underbrace{\sum_{T\subseteq S,\,|T|\geq 2,\,i\in T,\,K=\pi(i)\cap T} (-1)^{|S\setminus T|}}_{(\dagger)}\,. \end{aligned}$$

As $\hat{q}_D(i|\emptyset)=0$, to get the desired conclusion one needs to show, for any $\emptyset\neq K\subseteq\pi(i)$, that the internal sum $(\dagger)$ is 0 in case $i\neq j$ and $(-1)^{|R\setminus K|}$ if $i=j$. Indeed, if one denotes $W=S\setminus[\pi(i)\cup\{i\}]$, it can equivalently be written as

$$(\dagger) = \sum_{T=K\cup\{i\}\cup L,\,L\subseteq W} (-1)^{|S\setminus T|} = \sum_{L\subseteq W} (-1)^{|S\setminus(K\cup\{i\}\cup L)|} = \sum_{L\subseteq W} (-1)^{|(S\setminus(K\cup\{i\})\setminus L|}\,,$$

which vanishes in case $W\neq\emptyset$, and equals to $(-1)^{|S\setminus(K\cup\{i\})|}$ if $W=\emptyset$. Of course, $W=\emptyset$ if and only if $i=j$, in which case $S\setminus(K\cup\{i\})=R\setminus K$. □

6.2 Pure form of the first ILP task

Thus, the aim is to maximize the function

$$(\mathsf{a},\mathsf{c}) \longmapsto \sum_{S\subseteq N,\,|S|\geq 2} r_D^{\mathcal{Q}}(S)\cdot \mathsf{c}(S) \tag{25}$$

over the domain of vectors (a,c) with integer components satisfying the inequalities from § 3.2. By Theorem 5.1, this is equivalent to maximization of this function over the codes $(\mathsf{a}_H,\mathsf{c}_H)$ of chain graphs without flags equivalent to acyclic directed graphs. The function (25) only depends on the characteristic imset c_H, because $\mathsf{c}_H=\mathsf{c}_G$ for some (equivalent) acyclic directed graph G. By (19), the solution maximizes $\mathcal{Q}$ in the domain of acyclic directed graphs over N. An obvious modification is to replace (a.1) by (a.1*); see § 5.3.

7 Search space reduction

The idea of *pruning* components of the vector codes on the basis of a particular form of the database (and the criterion) is described in this section.

7.1 Pruning components of graph code

This is the idea elaborated in [6] and then applied in [10, 5]. The basic observation is the following, proved as Lemma 1 in [7], which is an extended version of [6]. We repeat the proof to pinpoint the core argument.

Lemma 7.1 *Let $\mathcal{Q}$ be an additively decomposable criterion and D a database such that, for some $i \in N$ and $B \subseteq N \setminus \{i\}$,*

$$\exists\, C \subset B \quad q_D(i|C) > q_D(i|B)\,. \tag{26}$$

Then no acyclic directed graph G over N with $\mathrm{pa}_G(i) = B$ maximizing $G \mapsto \mathcal{Q}(G, D)$ exists.

Proof: Assume for a contradiction that G is such a graph and consider the directed graph H obtained from G by the removal of arrows from nodes in $B \setminus C$ to i. Then, clearly, H is also an acyclic directed graph over N, and, by (5), one has

$$\mathcal{Q}(H, D) - \mathcal{Q}(G, D) = q_D(i \,|\, \mathrm{pa}_H(i)) - q_D(i \,|\, \mathrm{pa}_G(i)) = q_D(i|C) - q_D(i|B) > 0.$$

This contradicts the assumption that G maximizes $G \mapsto \mathcal{Q}(G, D)$. □

Lemma 7.1 means that, provided the condition (26) holds for some i and B, one cannot have an optimal acyclic directed graph G with $pa_G(i) = B$, for which reason, one can assume without loss of generality $\eta_G(i|B) = 0$ in (6). Thus, the component of η_G for $(i|B)$ and the respective local score can be excluded from the considerations in (6). This was the idea of *pruning* the components of η from [10], used as a pre-processing step there; it was mentioned there that, in practice, effective pruning can be done.

This is, in fact, based on further observations from [6]. For a general decomposable criterion $\mathcal{Q}$, the condition (26) only allows one to remove just one components of the η-vector. To remove most of them (= to reduce the exponential length in $|N|$ to the "polynomial" one) one has to verify exponentially many such conditions, which looks like the task of exponential complexity as well.

The point is that the criteria used in practice somehow give the preference to sparse graphs. For example, the BIC-criterion has a penalty term which, in fact, protects the large parent sets to occur in the optimal graph. Indeed, realize that, for $i \in N$, $B \subseteq N \setminus \{i\}$, the respective local contribution $\mathsf{dim}(i|B)$ to the dimension from (7) is exponential in $|B|$: $\prod_{j\in B} r(j) \geq 2^{|B|}$ (as $r(j) \geq 2$ for any $j \in N$). This often allows one to exclude in one step exponentially many parent sets (= components of the η-vector).

Lemma 7.2 *Let D be a database of the length $d \geq 2$, $i \in N$ and $C \subset B \subseteq N \setminus \{i\}$ with*

$$\prod_{j\in B} r(j) - \prod_{\ell\in C} r(\ell) > \frac{2}{\ln d \cdot \{r(i) - 1\}} \cdot \sum_{y\in \mathsf{X}_C} \sum_{z\in \mathsf{X}_i} d_{[y,z]} \cdot \ln \frac{d_{[y]}}{d_{[y,z]}}\,. \tag{27}$$

Then, there is no acyclic directed graph G over N maximizing $G \mapsto \mathsf{BIC}\,(G, D)$ with $\mathrm{pa}_G(i) = B$. In particular, the same conclusion holds if, for $i \in N$, $\emptyset \neq B \subseteq N \setminus \{i\}$,

$$\prod_{j\in B} r(j) - 1 > \frac{2 \cdot d}{\ln d \cdot \{r(i) - 1\}} \cdot \sum_{z\in \mathsf{X}_i} \frac{d_{[z]}}{d} \cdot \ln \frac{d}{d_{[z]}}\,. \tag{28}$$

Proof: The condition (28) is just (27) for $C = \emptyset$. The key argument, based on (7), is an upper estimate for $\mathsf{mll}_D(i|B) - \mathsf{mll}_D(i|C)$ derived from the well-known inequality $I(i; B \setminus C \,|\, C) \leq H(i|C)$, where $I(i; B \setminus C \,|\, C)$ is the conditional mutual information (with respect to the empirical probability measure induced by D) and $H(i|C)$ is the corresponding conditional entropy; see, for example, Lemma 3.1 in [16]:

$$\begin{aligned}
\mathsf{mll}_D(i|B) - \mathsf{mll}_D(i|C) &= \sum_{x\in \mathsf{X}_{B\setminus C}} \sum_{y\in \mathsf{X}_C} \sum_{z\in \mathsf{X}_i} d_{[x,y,z]}\cdot \ln\frac{d_{[x,y,z]}}{d_{[x,y]}} - \sum_{y\in \mathsf{X}_C} \sum_{z\in \mathsf{X}_i} d_{[y,z]}\cdot \ln\frac{d_{[y,z]}}{d_{[y]}} \\
&= \sum_{x\in \mathsf{X}_{B\setminus C}} \sum_{y\in \mathsf{X}_C} \sum_{z\in \mathsf{X}_i} d_{[x,y,z]}\cdot \ln\frac{d_{[x,y,z]}\cdot d_{[y]}}{d_{[x,y]}\cdot d_{[y,z]}} \\
&= d\cdot \sum_{x\in \mathsf{X}_{B\setminus C}} \sum_{y\in \mathsf{X}_C} \sum_{z\in \mathsf{X}_i} \{d_{[x,y,z]}/d\}\cdot \ln\frac{\{d_{[x,y,z]}/d\}\cdot \{d_{[y]}/d\}}{\{d_{[x,y]}/d\}\cdot \{d_{[y,z]}/d\}} \\
&\leq d\cdot \sum_{y\in \mathsf{X}_C} \sum_{z\in \mathsf{X}_i} \{d_{[y,z]}/d\}\cdot \ln\frac{\{d_{[y]}/d\}}{\{d_{[y,z]}/d\}} \\
&= \sum_{y\in \mathsf{X}_C} \sum_{z\in \mathsf{X}_i} d_{[y,z]}\cdot \ln\frac{d_{[y]}}{d_{[y,z]}}\,.
\end{aligned}$$

Thus, one can write by (27), because $\frac{\ln d}{2} > 0$ and $r(i)-1>0$,

$$\mathsf{mll}_D(i|B) - \mathsf{mll}_D(i|C) \leq \sum_{y\in \mathsf{X}_C} \sum_{z\in \mathsf{X}_i} d_{[y,z]}\cdot \ln\frac{d_{[y]}}{d_{[y,z]}} < \frac{\ln d}{2}\cdot \{r(i)-1\}\cdot \{\prod_{j\in B} r(j) - \prod_{\ell\in C} r(\ell)\}\,,$$

which, can be re-written, by (7), as follows:

$$\mathsf{bic}_D(i|B) = \mathsf{mll}_D(i|B) - \frac{\ln d}{2}\cdot \mathsf{dim}(i|B) < \mathsf{mll}_D(i|C) - \frac{\ln d}{2}\cdot \mathsf{dim}(i|C) = \mathsf{bic}_D(i|C)\,.$$

The rest follows from Lemma 7.1 applied to BIC. □

The inequality (27) has one important advantageous property: the right-hand side of (27) does not depend on B, while the left-hand side is a non-decreasing function of $|B|$. Therefore, once (27) holds for some $i\in N$ and $B\subseteq N\setminus\{i\}$ then (27) also holds for any B' with $B\subseteq B'\subseteq N\setminus\{i\}$. Hence, this allows us *in one step to prune exponentially many components* of η. Note that in [7] two such sufficient conditions for pruning with the BIC were derived and also one such a condition for the BDE (Theorem 2, 4 and 9 in [7]).

Because of the well-known upper estimate for the entropy term in the right-hand side of (28) one can derive from Lemma 7.2 the following observation, which is a minor strengthening of Theorem 2 from [7] (and of Theorem 1 from [6]).

Corollary 7.1 *Let D be a database of the length $d\geq 2$, $i\in N$ and $\emptyset\neq B\subseteq N\setminus\{i\}$ with*

$$\prod_{j\in B} r(j) - 1 > \frac{2\cdot d}{\ln d}\cdot \frac{\ln r(i)}{r(i)-1}\,. \tag{29}$$

Then, there is no acyclic directed graph G over N maximizing $G\mapsto \mathsf{BIC}\,(G,D)$ with $\mathrm{pa}_G(i)=B$. In particular, this is the case whenever $4\leq d<2^{|B|}-1$.

Proof: A well-known fact is that the entropy achieves its maximal value for the uniform distribution:

$$\sum_{z\in \mathsf{X}_i} \frac{d_{[z]}}{d}\cdot \ln\frac{d}{d_{[z]}} \leq \sum_{z\in \mathsf{X}_i} \frac{1}{r(i)}\cdot \ln\frac{r(i)}{1} = \ln r(i)\,,$$

which means that (29) implies (28). As concerns the last claim, realize that $d \geq 4$ implies $\frac{\ln d}{2} \geq \frac{\ln 4}{2} = \ln 2 > 0$ and, hence, $\frac{1}{\ln 2} \geq \frac{2}{\ln d} > 0$. As $r(i) \geq 2$ and the function $x \mapsto \frac{x-1}{\ln x}$ is increasing on $(1, \infty)$, one has $\frac{r(i)-1}{\ln r(i)} \geq \frac{2-1}{\ln 2} = \frac{1}{\ln 2} > 0$ and, hence, $\ln 2 \geq \frac{\ln r(i)}{r(i)-1} > 0$. By multiplying these inequalities we get, using the assumption $d < 2^{|B|} - 1$,

$$1 = \frac{1}{\ln 2} \cdot \ln 2 \geq \frac{2}{\ln d} \cdot \frac{\ln r(i)}{r(i)-1} > 0 \quad \Rightarrow \quad \prod_{j \in B} r(j) - 1 \geq 2^{|B|} - 1 > d \geq d \cdot \frac{2}{\ln d} \cdot \frac{\ln r(i)}{r(i)-1},$$

which means (29) holds, no matter what the database D is. □

The meaning of the last claim in Corollary 7.1 is that if the database is not lengthy enough, than one can always exclude some large parent sets. The above described procedure, therefore, allows one to prune all components $\eta(i|B)$ with $|B|$ high enough.

Actually, as reported in § 6 of [7], the pruning procedure was applied to some databases from the so-called UCI repository, and it typically resulted in the reduction of the parent set cardinality to at most 5; only in a few cases the maximal parent set cardinality was 7 or 8. The reduction was done both for the BIC-score and the BDE-score. However, no particular algorithm or a formal procedure to perform the pruning was presented in [7]; it was only mentioned there that the above mentioned principles were applied. Also, as one can deduce from the laconic notes from [7], the pruning was probably quite costly procedure from the point of view of computational time.

Remark Coming back to Lemma 7.2: note that one can derive the same conclusion there if (27) is replaced by the condition

$$\prod_{j \in B} r(j) - \prod_{\ell \in C} r(\ell) > \frac{2}{\ln d \cdot \{r(i)-1\}} \cdot \sum_{y \in \mathsf{X}_C} \sum_{x \in \mathsf{X}_{B \setminus C}} d_{[y,x]} \cdot \ln \frac{d_{[y]}}{d_{[y,x]}}.$$

This is based on an alternative estimate $I(i; B \setminus C \,|\, C) \leq H(B \setminus C \,|\, C)$. However, the expression on the right-hand side of the inequality *does depend* on B and may be difficult to compute. Therefore, the inequality may be difficult to verify. On the other hand, this alternative condition may be valid even if (27) is not and may justify pruning which cannot be derived using (27).

7.2 Pruning components of the characteristic imset

The point is that the pruning of the η-vector can be utilized in this context.

Corollary 7.2 *Let $\mathcal{Q}$ be a decomposable criterion and D a database such that, for some $S \subseteq N$, $|S| \geq 2$, the condition (26) holds for any $i \in S$ and B with $S \setminus \{i\} \subseteq B \subseteq N \setminus \{i\}$. Then $\mathsf{c}_G(S) = 0$ for any acyclic directed graph G over N maximizing $G \mapsto \mathcal{Q}(G, D)$.*

Proof: Assume for a contradiction that an optimal acyclic directed graph G with $\mathsf{c}_G(S) = 1$ exists. Then, by (4), there exists $i \in S$ and $B \subseteq N \setminus \{i\}$ with $S \setminus \{i\} \subseteq B$ such that $\eta_G(i|B) = 1$, which means $pa_G(i) = B$. However, since (26) holds for D and the pair $(i|B)$, by Lemma 7.1, G is not optimal, which contradicts the assumption. □

This motivated the following definition.

Definition 7.1 Let $\mathcal{Q}$ be a decomposable criterion and D a database. Given a pair $(i|B)$, where $i \in N$ and $B \subseteq N \setminus \{i\}$, we say that the respective component $\eta(i|B)$ (of the η-vector) is *redundant* (for D and $\mathcal{Q}$) if the condition (26) holds. Analogously, given $S \subseteq N$, $|S| \geq 2$, we say that the corresponding component $\mathsf{c}(S)$ of the c-vector is *redundant* (for D and $\mathcal{Q}$) if, for every $i \in S$ and every $S \setminus \{i\} \subseteq B \subseteq N \setminus \{i\}$, the components $\eta(i|B)$ is redundant.

Indeed, the redundant components of c can be ignored if our task is to find an optimal BN structure. By Corollary 7.2, if $\mathsf{c}(S)$ is redundant, one always has $\mathsf{c}_G(S) = 0$ in any optimal graph G. In other words, in the task (25) one can assume without loss of generality $\mathsf{c}(S) = 0$ and the value $r_D^{\mathcal{Q}}(S)$ is not needed to solve such a simplified ILP task. Simply, one can concentrate on the *non-redundant* components of c only.

For our purpose, it would be convenient to have an algorithm for the removal of all redundant components of c, because this seems to lead to the maximal reduction in the dimension of the corresponding ILP problem. However, as indicated in § 7.1, such an algorithm may appear to be too complex, if its goal is really to indicate all redundant components. Then the question is whether the computational time spent on finding such a maximal reduction would be worthwhile.

Therefore, what is suggested below instead is a heuristic procedure, whose aim is to find inclusion-maximal non-redundant components. Of course, this preliminary version of the procedure can probably be modified later to become more efficient. Its goal is to get a system of sets $\mathcal{T} \subseteq \{S \subseteq N;\ |S| \geq 2\}$ which is

- closed under subsets: $S \in \mathcal{T}$, $T \subseteq S$, $|T| \geq 2$ implies $T \in \mathcal{T}$,
- all “non-redundant” sets are included in it: if $\mathsf{c}(S)$ is non-redundant for some $S \subseteq N$, $|S| \geq 2$ then $S \in \mathcal{T}$,
- $\mathcal{T}$ consists of sets of small cardinality: there exists “low” upper bound $t \in \mathbb{N}$ such that $|S| \leq t$ for any $S \in \mathcal{T}$.

The practical experiments described in § 6 of [7] suggest that this is a realistic plan. Thus, the goal of the procedure is to obtain inclusion-maximal sets S for which $\mathsf{c}(S)$ is non-redundant. Let us consider $S \subseteq N$, $|S| \geq 2$ such that $\mathsf{c}(T)$ is redundant for any its strict subset T, $S \subset T \subseteq N$. If this is the case, then $\eta(i|B')$ is redundant whenever $S \setminus \{i\} \subset B' \subseteq N \setminus \{i\}$ (see Definition 7.1). Thus, $\mathsf{c}(S)$ is redundant iff, for every $i \in S$, $\eta(i \,|\, S \setminus \{i\})$ is redundant (= the condition (26) holds for any $i \in S$ and $B = S \setminus \{i\}$). Thus, the (inclusion-maximal) non-redundance means that there exists $i \in S$ with (inclusion-maximal) non-redundant $\eta(i \,|\, S \setminus \{i\})$. This, perhaps, explains why the following procedure makes sense.

Heuristic procedure

Let D be a database and $\mathcal{Q}$ a score equivalent and additively decomposable criterion.

1. The first step is to derive a suitable formula for local scores. Because, in (26), one can replace $q_D(*|*)$ by its standardized version $\hat{q}_D(*|*)$ (uniquely determined by $\mathcal{Q}$, it is often convenient to have a subroutine which allows one to compute any standardized local score $\hat{q}_D(i|B)$ whenever it is needed. (One of the possible ways is to have a formula for the components of the data vector $t_D^{\mathcal{Q}}(S)$ for $S \subseteq N$, $|S| \geq 2$ and use (23) for our purpose.)

2. The second step is, for any fixed $i \in N$, to find all inclusion-maximal $B \subseteq N \setminus \{i\}$ such that $\eta(i|B)$ is non-redundant (in sense of Definition 7.1).

   - First, we need to indicate as redundant those $\eta(i|B)$ in which $|B|$ is high enough. In the case of the BIC-score, one can use the condition (28) from Lemma 7.2 for this purpose. (In the case of the BDE-score one needs to have an analogous sufficient condition like in Theorem 9 of [7].)
   - Second, if we already know that $|B|$ has some (not necessarily tight) upper bound k, then we can inductively, for $\ell = 0, \dots k$, indicate the inclusion-maximal B with $\eta(i|B)$ non-redundant and $|B| \leq \ell$. We already know that $\eta(i|\emptyset)$ is non-redundant; therefore, the starting list for $\ell = 0$ contains just $B = \emptyset$. Now, for $\ell = 1, \dots, k$, we compute (inductively) for any $B \subseteq N \setminus \{i\}$ with $|B| = \ell$ both the value of $q(i|B)$ and the value of $q^*(i|B) \equiv \max\{\, q(i|C);\ C \subset B\}$. Clearly, $\eta(i|B)$ is redundant iff $q(i|B) < q^*(i|B)$. If it is non-redundant, we include B in the list and remove from the list all $C \subset B$, which have been there before. Observe that $q^*(i|B)$ can be computed inductively.

   Let $\mathcal{C}$ denote the resulting collection of pairs $(i|B)$.

3. The third step is to define, on the basis of the (obtained) collection $\mathcal{C}$ of pairs $(i|B)$ with the property that $\eta(i|C)$ is redundant provided there is no $(i|B) \in \mathcal{C}$ with $C \subseteq B$:

$$\mathcal{T} := \{S \subseteq N\,;\ |S| \geq 2\ \ \exists\,(i|B) \in \mathcal{C}\ \text{ such that } S \subseteq \{i\} \cup B\,\}\,.$$

   Of course, $\mathcal{T}$ contains all sets S with non-redundant $\mathsf{c}(S)$. If $\mathcal{C}$ is the class from step 2., the maximal sets in $\mathcal{T}$ are just the maximal sets S^* with non-redundant $\mathsf{c}(S^*)$.

4. We establish a cache for values that correspond to sets $S \in \mathcal{T}$ and use (24) in Corollary 6.1 to compute the values $r_D^{\mathcal{Q}}(S)$ for $S \in \mathcal{T}$. The goal of our later optimization task will be, in fact, to maximize $\mathsf{c} \mapsto \sum_{S \in \mathcal{T}} r_D^{\mathcal{Q}}(S) \cdot \mathsf{c}(S)$. (An alternative method is to compute first the values $t_D^{\mathcal{Q}}(S)$ for $S \in \mathcal{T}$, and the use (22) to get what is needed.)

Another option is to start with a store of local score values $q(*|*)$ which has already been pruned by somebody else, by which we mean that the local scores not involved in the store correspond to redundant components of η. In that case, one can use Corollary 6.1; but, what is needed here is to compute $q_D(i|D)$ whenever $D \subseteq B$ and $\eta(i|B)$ is non-redundant.

8 Second phase: reconstruction of the essential graph

Assume that the first ILP problem has been successfully solved and a solution (a, c) found, which is a vector code for a chain graph H without flags equivalent to an acyclic directed graph over N. The graph H need not be an essential graph, but we are already sure that the vector c is the characteristic imset for some acyclic directed graph G over N, that is, $\mathsf{c} = \mathsf{c}_G$.

The idea is to fix c now and search through all a' such that $(\mathsf{a}', \mathsf{c})$ satisfies the inequalities from § 3.2. This, in fact, means to search through all codes of chain graphs without flags equivalent to G. By Lemma 2.1, the class $\mathcal{H}$ of these graphs has the largest graph, namely the essential graph G^* for the equivalence class of acyclic directed graphs $\mathcal{G}$ containing G.

The largest graph in $\mathcal{H}$ can be characterized as the graph with the minimal amount of arrowheads in $\mathcal{H}$. Thus, it minimizes the function

$$H \in \mathcal{H} \;\longmapsto\; \sum_{i,j\in N,\, i\neq j} \mathsf{a}_H(i \to j)\,.$$

In this (second) ILP problem, since c is fixed, then number of constrains in the LP relaxation from § 3.2 is already polynomial in $|N|$. We only need the values $\mathsf{c}(S)$ for $2 \leq |S| \leq 3$ and use the inequalities (b.1), (c.1) and (c.3)-(c.5). The number of variables $\mathsf{a}(i \to j)$ is also polynomial in $|N|$.

If someone is particularly interested in finding an acyclic directed graph as a solution (which, however, need not be unique), then one can simply maximize the above-mentioned function instead. This is because the elements of $\mathcal{G}$ (= Markov equivalence class of acyclic directed graphs equivalent to G) can be viewed as the elements in $\mathcal{H}$ with the maximal amount of arrowheads. Actually, one can even ascribe different positive weights $w(a \to j)$ to possible arrowheads and, determine in that way a criterion for the choice an acyclic directed graph from $\mathcal{G}$, which takes into account our preference od directions of arrows. Then the aim should be to maximize the function

$$H \in \mathcal{H} \;\longmapsto\; \sum_{i,j\in N,\, i\neq j} \mathsf{a}_H(i \to j)\cdot w(i \to j)\,.$$

The other option is to include the second ILP task in the first one, that is, to consider the following maximization task over integral vectors (a,c) satisfying the inequalities from § 3.2:

$$(\mathsf{a},\mathsf{c}) \;\mapsto \max\left[\sum_{S\subseteq N,\,|S|\geq 2} r_D^{\mathcal{Q}}(S)\cdot \mathsf{c}(S) - \varepsilon\cdot \sum_{i,j\in N,\, i\neq j} \mathsf{a}(i\to j)\right],$$

where $r_D^{\mathcal{Q}}$ is the respective *revised data vector* and $\varepsilon > 0$ small enough in comparison with the absolute values of the components of $r_D^{\mathcal{Q}}$, so that the c-part of the solution is ensured to maximize $\mathsf{c} \mapsto \sum_{S\subseteq N,\,|S|\geq 2} r_D^{\mathcal{Q}}(S)\cdot \mathsf{c}(S)$. The technical problem here could be how small ε is allowed to avoid rounding errors.

9 Complexity considerations

After the search space reduction we have a class of sets $\mathcal{T} \subseteq \{S \subseteq N;\ |S| \geq 2\}$ which is closed under subsets and there is a reasonable upper bound $t \in \mathbb{N}$ on the cardinality of sets in $\mathcal{T}$: $|S| \leq t$ for $S \in \mathcal{T}$. The hope is that typically t will be 5 or 6. The aim is to maximize the function

$$\mathsf{c}_G \mapsto \sum_{S\subseteq N,\,|S|\geq 2} r_D^{\mathcal{Q}}(S)\cdot \mathsf{c}_G(S),$$

which is, by (19), equivalent to the maximization of $\mathcal{Q}(G,D)$ over the class of characteristic imsets c_G of acyclic directed graphs G over N, such that, moreover, one has

$$\mathsf{c}_G(S) = 0 \qquad \text{for} \quad S \subseteq N,\ S \not\in \mathcal{T}.$$

The result of the pre-processing stage should be a store of values for $r_D^{\mathcal{Q}}(S)$ for $S \in \mathcal{T}$. This is also the situation when we deal with the problem of restricted learning with prescribed upper bound $t \in \mathbb{N}$ for the size of cliques in the moral graph of G.

In such a situation, most of the inequalities (e.1)-(e.2) from § 3.2 can be omitted, as they involve zero values and are trivially valid. It is enough to consider the following inequalities:

(e.1$\star$) $\forall\, S\subseteq N,\ |S|\geq 3,\ S\in\mathcal{T}^{\star} \qquad \sum_{i\in S}\mathsf{c}(S\setminus\{i\})\leq 2+(|S|-2)\cdot\mathsf{c}(S)$,

(e.2$\star$) $\forall\, S\subseteq N,\ |S|\geq 4,\ S\in\mathcal{T} \qquad (|S|-1)\cdot\mathsf{c}(S)\leq\sum_{i\in S}\mathsf{c}(S\setminus\{i\})$,

with $\mathcal{T}^{\star}=\{S\subseteq N;\ \exists\, i\in S\quad S\setminus\{i\}\in\mathcal{T}\}$. We implicitly assume that $t\equiv\max\{|S|;\ S\in\mathcal{T}\}$ is small and fixed; thus, the number of inequalities (e.1)-(e.2) is, in fact, reduced polynomially, to at most $\mathcal{O}(|N|^{t+1})$ size.

Unfortunately, this is not the case with the inequalities (a.1), respectively (a.1*). One can always have a chordless cycle of arbitrary high length. In particular, there is no hope to reduce the number of inequalities in (a.1) to a polynomial number in $|N|$, even in the case of restricted learning. In case of acyclicity inequalities, however, each inequality has quite clear interpretation: the inequality for $S\subseteq N$, $|S|\geq 4$ means a forbidden chordless (either undirected or semi-directed) cycle composed just of the nodes in S in the respective graph H. Thus, an iterative constraint adding procedure is suggested below.

9.1 Summary of the procedure

1. The aim should be to maximize the function

$$(\mathsf{a},\mathsf{c})\quad\longmapsto\sum_{S\subseteq N,|S|\geq 2} r_{D}^{\mathcal{Q}}(S)\cdot\mathsf{c}(S)=\sum_{S\in\mathcal{T}} r_{D}^{\mathcal{Q}}(S)\cdot\mathsf{c}(S)$$

over the lattice points within the polyhedron specified by the inequalities from § 3.2 such that $\mathsf{c}(S)=0$ for $S\notin\mathcal{T}$. We know, by Theorem 5.1, that the c-parts of these vectors are nothing but the characteristic imsets for acyclic directed graphs over N.

2. As the first iteration, we relax this ILP problem on the basis of $\mathcal{T}$ and consider only the inequalities (b.1)-(b.2), (c.1)-(c.5), (e.1$\star$)-(e.2$\star$). The number of these inequalities should be reasonable, of at most $\mathcal{O}(|N|^{t+1})$ size, where $t=\max\{|S|;\ S\in\mathcal{T}\}$.

3. The solution of the starting ILP problem should be code $(\mathsf{a}_H,\mathsf{c}_H)$ of a hybrid 3-acyclic graph H without flags, by Lemma 5.3 (i).

4. We check whether H is a chain graph equivalent to an acyclic directed graph by searching for chordless semi-directed or undirected cycles in H. This should be polynomially complex graphical procedure.

   - If there is no such a cycle, the graph H encodes the desired solution.
   - If not, we find all such chordless cycles in H, take the respective acyclicity inequalities (a.1), respectively (a.1*), and incorporate them in the considered system of inequalities and run the new ILP problem (= come to step 2. with an extended class of inequalities).

5. Provided we do not get stuck with memory overload, the solution will be a vector $(\mathsf{a}_H,\mathsf{c}_H)$, where H is a chain graph over N, which has no flag and is equivalent to an acyclic directed graph. We find the corresponding essential graph by establishing the second ILP problem, as described in § 8.

A modified version of the above procedure is that in Step 2., we already include the inequalities (a.1) for $S \in \mathcal{T}$. Then, in step 3., if a chordless cycle is found, it should have as the node set some $T \subseteq N$ not in $\mathcal{T}$.

References

[1] S. A. Andersson, D. Madigan, M. D. Perlman: A characterization of Markov equivalence classes for acyclic digraphs, Annals of Statistics 25 (1997) 505-541.

[2] R. R. Bouckaert: Bayesian belief networks - from construction to evidence, PhD thesis, University of Utrecht 1995.

[3] D. M. Chickering: Optimal structure identification with greedy search, Journal of Machine Learning Research 3 (2002) 507-554.

[4] J. Cussens: Maximum likelihood pedigree reconstruction using integer programming, in Proceeedings of the Workshop on Constraint Based Methods for Bioinformatics (WCBMB) 2010, pp. 9-19.

[5] J. Cussens: Bayesian network learning with cutting planes, in Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI) 2011, 153-160.

[6] C. P. de Campos, Z. Zeng, Q. Ji: Structure learning Bayesian networks using constraints, in Proceedings of the 26th International Conference on Machine Learning (ICML) 2009, pp. 113-120.

[7] C. P. de Campos, Q. Ji: Efficient structure learning Bayesian networks using constraints, Journal of Machine Learning Research 12 (2011) 663-689.

[8] D. Heckerman, D. Geiger, D. M. Chickering: Learing Bayesian networks - the combination of knowledge and statistical data, Machine Learning 20 (1995) 194-243.

[9] R. Hemmecke, S. Lindner, M. Studený: Characteristic imsets for learning Bayesian network structure, to appear in International Journal of Approximate Reasoning (2012); see `doi:10.1016/j.ijar.2012.04.001`.

[10] T. Jaakkola, D. Sontag, A. Globerson, M. Meila: Lerning Bayesian network structure using LP relaxations, in Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, pp. 358-365.

[11] S. L. Lauritzen: Graphical Models, Clarendon Press 1996.

[12] S. Lindner: Discrete optimization in machine learning - learning Bayesian network structures and conditional independence implication, PhD thesis, TU Munich 2012.

[13] R. E. Neapolitan: Learning Bayesian Networks, Pearson Prentice Hall 2004.

[14] G. E. Schwarz: Estimation of the dimension of a model, Annals of Statistics 6 (1978) 461-464.

[15] M. Studený: A recovery algorithm for chain graphs, International Journal of Approximate Reasoning 17 (1997) 265-293.

[16] M. Studený, J. Vejnarová: The multiinformation function as a tool for measuring stochastic dependence, in Learning in Graphical Models (M. I. Jordan ed.), Kluwer 1998, pp. 261-298.

[17] M. Studený: Characterization of essential graphs by means of the operation of legal merging of components, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 12 (2004) 43-62.

[18] M. Studený: Probabilistic Conditional Independence Structures, Springer 2005.

[19] M. Studený, J. Vomlel, R. Hemmecke: A geometric view on learning Bayesian network structures, International Journal of Approximate Reasoning 51 (2010) 578-586.

[20] M. Studený, R. Hemmecke, S. Lindner: Characteristic imset - a simple algebraic representative of a Bayesian network structure, in Proceedings of the 5th European Workshop on Probabilistic Graphical Models (PGM) 2010, pp. 257-264.

[21] M. Studený, D. Haws: On polyhedral approximations of polytopes for learning Bayes nets, research report n. 2303, Institute of Information Theory and Automation of the ASCR, Prague, July 2011, also available on `http://arxiv.org/abs/1107.4708`.

[22] M. Studený, D. Haws, R. Hemmecke, S. Lindner: Polyhedral approach to statistical learning graphical models, in Proceedings of the 2nd CREST–SBM International Conference, World Scientific 2012, pp. 346-372.

[23] T. Verma, J. Pearl: Equivalence and synthesis of causal models, in Proceedings of the 6th conference on Uncertainty in Artificial Intelligence, Elsevier 1991, pp. 220-227.