Parent Assignment Is Hard for the MDL, AIC, and NML Costs

Mikko Koivisto

HIIT Basic Research Unit, Department of Computer Science
Gustaf Hällströmin katu 2b, FIN-00014 University of Helsinki, Finland
mikko.koivisto@cs.helsinki.fi

Abstract. Several hardness results are presented for the parent assignment problem: Given m observations of n attributes $x_1, \ldots, x_n$, find the best parents for x_n, that is, a subset of the preceding attributes so as to minimize a fixed cost function. This attribute or feature selection task plays an important role, e.g., in structure learning in Bayesian networks, yet little is known about its computational complexity. In this paper we prove that, under the commonly adopted full-multinomial likelihood model, the MDL, BIC, or AIC cost cannot be approximated in polynomial time to a ratio less than 2 unless there exists a polynomial-time algorithm for determining whether a directed graph with n nodes has a dominating set of size $\log n$, a LOGSNP-complete problem for which no polynomial-time algorithm is known; as we also show, it is unlikely that these penalized maximum likelihood costs can be approximated to within any constant ratio. For the NML (normalized maximum likelihood) cost we prove an NP-completeness result. These results both justify the application of existing methods and motivate research on heuristic and super-polynomial-time algorithms.

1 Introduction

Structure learning in Bayesian networks is often approached by minimizing a sum of costs assigned to each local structure consisting of an attribute and its parents [1, 2]. If an ordering of the attributes is given, then the subtasks of assigning optimal parents to each attribute can be solved independently of each other. Unfortunately, for many objective functions of interest, no polynomial time algorithm is known, unless one is willing to bound the number of parents above by a constant, in which case the problem can be solved in polynomial time. Consequently, researchers have proposed greedy algorithms with no performance guarantees [1] and heuristic branch-and-bound methods that find a global optimum but can be very slow in the worst case [3, 4]. However, the precise complexity of the parent assignment problem, even for the most commonly used cost functions, is unknown.

This paper focuses on the following variant of the parent assignment problem: given a data set containing m observations on n discrete attributes $x_1, \ldots, x_n$, find the parents $x_{s_1}, \ldots, x_{s_k}$ for x_n so as to minimize the Minimum Description

G. Lugosi and H.U. Simon (Eds.): COLT 2006, LNAI 4005, pp. 289–303, 2006.

Length (MDL) cost [5] under the full-multinomial model[1] [6]. This commonly adopted cost function has the important property that the optimal number of parents is always at most $\log m$ (throughout this paper we write $\log m$ for $\lceil \log_2 m \rceil$). This is because the number of parameters of the multinomial model grows exponentially in the number of parents k, whereas the error, or the negative log-likelihood, grows at most linearly in the number of observations m [7]. That said, only $O(n^{\log m})$ smallest subsets of the $n-1$ attributes need to be evaluated, suggesting that the problem is very unlikely to be NP-hard (see, e.g., Papadimitriou and Yannakakis [8]). Still, it is important and intriguing to determine whether the problem can be solved in polynomial time.

In this paper we show that for the (two-part) MDL cost the parent assignment problem is LOGSNP-hard, in other words, at least as hard as determining whether a directed graph with n nodes has a dominating set of size $\log n$ [8]; for this Log Dominating Set problem no polynomial-time algorithm is known.

Having this result, it is natural to ask whether similar results hold for other penalized maximum likelihood costs, such as Akaike's information criterion (AIC) [9] and the Normalized Maximum Likelihood (NML) criterion [10, 11]; note that the Bayesian information criterion (BIC) [12] coincides with the MDL cost. Our finding is that while MDL and AIC obey identical characterizations in terms of LOGSNP-hardness, the behavior of NML seems to be radically different: On one hand, we show that approximating the MDL or AIC cost to a ratio less than 2 is LOGSNP-hard. On the other hand, for the NML cost we can obtain an NP-completeness result; however, we currently do not know any nontrivial inapproximability result for NML.

While these results are somewhat theoretical and perhaps not very surprising, they provide evidence that the considered parent assignment problem is very unlikely to have a polynomial-time algorithm with a good quality guarantee. This justifies and motivates the application of existing search heuristics and, more importantly, research on novel super-polynomial-time algorithms.

The rest of this paper is structured as follows. In Sect. 2 we formulate some decision and optimization variants of the parent assignment problem for penalized maximum likelihood costs under the full-multinomial model. In Sect. 3 we prove the LOGSNP-hardness result for MDL by a simple reduction from Log Dominating Set; this part introduces the reduction in a relatively easy and clean manner. We then use essentially the same reduction in Sect. 4 to prove the inapproximability results for MDL and AIC. We consider the case of NML in Sect. 5. In Sect. 6 we discuss some open problems and related previous work.

2 Preliminaries

For simplicity, we restrict our consideration to $\{0,1\}$-valued attributes. Let X be an $m \times n$ data matrix, where the entry at the ith row and jth column, denoted as x^i_j, represents the ith observation of the jth attribute; submatrices are referred

[1] In the full-multinomial model each value configuration of the parent attributes is assigned an independent multinomial distribution of x_n.

to by indexing with subsets of row and column indexes. To distinguish between attributes and columns of the data matrix we denote x_j and x_S for the attributes, but $\mathbf{x}_j$ and $\mathbf{x}_S$ for the respective columns of X.

2.1 Penalized Maximum Likelihood Under the Full-Multinomial Model

The multinomial model of conditional probability concerns the probability distribution of a "child" variable, say x_n^i, given a set of "parents," say x_S^i where $S \subseteq \{1, \ldots, n-1\}$. In case of binary variables, this model has $2^{|S|}$ parameters $\theta_{1|u}$, one for each possible value u of x_S, specifying the probability of $x_n^i = 1$ given $x_S^i = u$, for all $i = 1, \ldots, m$; this is, in fact, a Bernoulli distribution for each value of x_S. It is convenient to also define $\theta_{0|u} = 1 - \theta_{1|u}$. The m observations are treated as independent draws, so that the total likelihood of $\mathbf{x}_n$, conditionally on $\mathbf{x}_S$, is given by

$$\prod_{i=1}^{m} \theta_{x_n^i | x_S^i} = \prod_{u \in \{0,1\}^{|S|}} \prod_{v \in \{0,1\}} \theta_{v|u}^{m_{uv}},$$

where $m_{uv} = |\{i : x_S^i = u, x_n^i = v\}|$ is the number of observations that has value u on columns S and value v on column n. It is easy to find the maximizing parameter values: $\theta_{v|u} = m_{uv}/m_u$, where $m_u = |\{i : x_S^i = u\}|$ is the number of observations that has value u on column n.

Various forms of penalized maximum likelihood can be used as a criterion for choosing between different sets of parents. These criteria operate quantitatively in the logarithmic scale. The negative of the maximum log likelihood,

$$\beta(\mathbf{x}_S, \mathbf{x}_n) = -\sum_{u \in \{0,1\}^{|S|}} \sum_{v \in \{0,1\}} m_{uv} \log \frac{m_{uv}}{m_u},$$

gets a small value when the model fits well the data; $\beta(\mathbf{x}_n, \mathbf{x}_S)$ can be viewed as the number of bits needed to describe $\mathbf{x}_n$ given $\mathbf{x}_S$ and the estimated model parameters. The MDL, AIC, and NML criteria introduce specific additive penalization terms α_{MDL}, α_{AIC}, and α_{NML}, respectively, defined by

$$\begin{aligned} \alpha_{\mathrm{MDL}}(X, S) &= 2^{|S|-1} \log m\,, \\ \alpha_{\mathrm{AIC}}(X, S) &= 2^{|S|}\,, \\ \alpha_{\mathrm{NML}}(X, S) &= \log \sum_{\mathbf{x}'_n \in \{0,1\}^m} 2^{-\beta(\mathbf{x}_S, \mathbf{x}'_n)}\,. \end{aligned}$$

As $\alpha_{\mathrm{MDL}}(X, S)$ and $\alpha_{\mathrm{AIC}}(X, S)$ depend on (X, S) only through the number of rows m in X and the number of elements in S, we may conveniently treat them as functions of $(m, |S|)$. If M is a label of a criterion, e.g., from {MDL, AIC, NML}, we define the corresponding penalized maximum likelihood cost as

$$\gamma_M(X, S) = \alpha_M(X, S) + \beta(\mathbf{x}_S, \mathbf{x}_n)\,.$$

Notice that $2^{-\gamma_{\mathrm{NML}}(X,S)}$ is a conditional probability distribution of $\mathbf{x}_n$ given $\mathbf{x}_S$.

2.2 Variants of the Parent Assignment Problem

We will formally look at the parent assignment problem in the guise of one optimization problem as well as of two decision problems, which are suitable for complexity considerations. The following problems will be fixed once the penalty term α has been fixed:

Min Parent Assignment (α)
Input: A 0-1 matrix X of size $m \times n$.
Output: A subset $S \subseteq \{1, \ldots, n-1\}$ such that $\gamma(X, S) = \alpha(X, S) + \beta(\mathbf{x}_S, \mathbf{x}_n)$ is minimized.

Parent Assignment (α)
Instance: A 0-1 matrix X of size $m \times n$ and a number t.
Question: Is there a subset $S \subseteq \{1, \ldots, n-1\}$ such that $\gamma(X, S) = \alpha(X, S) + \beta(\mathbf{x}_S, \mathbf{x}_n)$ is at most t?

Small Parent Assignment (α)
Instance: A 0-1 matrix X of size $m \times n$ and numbers t and k.
Question: Is there a subset $S \subseteq \{1, \ldots, n-1\}$ of size at most k such that $\gamma(X, S) = \alpha(X, S) + \beta(\mathbf{x}_S, \mathbf{x}_n)$ is at most t?

Of these problems, Min Parent Assignment is the most natural optimization formulation of the parent assignment problem. Obviously, it is at least as hard as the corresponding decision variant, Parent Assignment. The second decision problem, Small Parent Assignment, involves an upper bound for the number of parents, which renders it at least as hard as Parent Assignment; we will not consider Small Parent Assignment until in Sect. 5.

3 MDL Parent Assignment Is Hard

In this section we show that Parent Assignment (α_{MDL}), or MDL-PA for short, is LOGSNP-hard. Papadimitriou and Yannakakis [8] defined the complexity class LOGSNP in order to capture computational problems that are unlikely to be NP-hard but very likely to have time complexity that scales, roughly, as $n^{\log n}$ where n is the input size.

Our proof is based on a reduction from a restricted dominating set problem defined below. As usual, for a directed graph G we call a node subset S a *dominating set* if each node i outside S is *dominated* by some node j in S, i.e., (i, j) is an arc in G.

Log Dominating Set (Log-DS)
Instance: A directed graph with $n-1$ nodes.
Question: Does the graph have a dominating set of size $\log n$?

A couple of details are here worth noting. First, we define the problem in terms of $n-1$ rather than n nodes, as this leads to somewhat simpler expressions in

the sequel. Second, the standard problem definition (e.g., Papadimitriou and Yannakakis [8]) has $\log n$ replaced by $\log(n-1)$ (or $n-1$ by n), however, it is not difficult to show that the two problems are polynomially equivalent.

We known that Log Dominating Set is an ideal representative of the class LOGSNP:

Theorem 1 ([8]). Log Dominating Set *is LOGSNP-complete.*

A key observation we will exploit in our reduction is that the maximum likelihood score is highly sensitive to "collisions." We say that a subset S *has a collision* in X if there exist two rows i and i' such that

$$x^i_S = x^{i'}_S \quad \text{but} \quad x^i_n \neq x^{i'}_n .$$

Thus, a collision occurs if some value on the parents appears with both values, 0 and 1, on the child.

On one hand, if no collision occurs, then the fit is perfect.

Lemma 1. *Let X be a* 0-1 *matrix of size $n \times n$ and S a subset of $\{1, \ldots, n-1\}$. If S has no collision in X, then $\beta(\mathbf{x}_n, \mathbf{x}_S) = 0$.*

Proof. Suppose that S has no collision in X. Then for any u either $m_{u0} = 0$ or $m_{u1} = 0$ or both. Thus, either $m_{u0} = m_u$ or $m_{u1} = m_u$, implying $m_{u0} \log(m_{u0}/m_u) + m_{u1} \log(m_{u1}/m_u) = m_u \log 1 = 0$. As $\beta(\mathbf{x}_n, \mathbf{x}_S)$ is a sum of these terms, one for each value of u, it must equal 0. □

On the other hand, the more collisions, the larger the minimum error. We will use the following lower bound.

Lemma 2. *Let X be a* 0-1 *matrix of size $n \times n$ and S a subset of $\{1, \ldots, n-1\}$. If S has a collision in X, then $\beta(\mathbf{x}_n, \mathbf{x}_S) \geq 2$.*

Proof. Suppose that $x^i_S = x^{i'}_S = u$ and $0 = x^i_n \neq x^{i'}_n = 1$. Since $m_{u1}, m_{u0} \geq 1$ and $m_{u1} + m_{u0} = m_u$, we have

$$\beta(\mathbf{x}_n, \mathbf{x}_S) \geq -\Big(m_{u0} \log \frac{m_{u0}}{m_u} + m_{u1} \log \frac{m_{u1}}{m_u} \Big) \geq \min_{0<p<1} \{ -\log p - \log(1-p) \} = 2 .$$

□

To amplify the effect of a collision, we consider simple repetitions. We say that an $m \times n$ matrix B is obtained by stacking r copies of a $q \times n$ matrix A, or, that B is the *r-stack* of A, if $m = rq$ and the $(tq + i)$th row vector of B equals the ith row vector of A for all $t = 0, \ldots, r-1$ and $i = 1, \ldots, q$.

Lemma 3. *Let X be a* 0-1 *matrix of size $n \times n$ and S a subset of $\{1, \ldots, n-1\}$. Let X' be the matrix obtained by stacking r copies of X. Then $\beta(\mathbf{x}'_n, \mathbf{x}'_S) = r \cdot \beta(\mathbf{x}_n, \mathbf{x}_S)$.*

Proof. For X' the maximum likelihood estimate for any parameter $\theta_{v|u}$ is simply $(rm_{uv})/(rm_u) = m_{uv}/m_u$, that is, the same as for the original matrix X. □

We apply these simple observations with the following strategy. First, we map an arbitrary instance of Log-DS, a graph G with $n-1$ nodes, to a suitable square matrix X of size $n \times n$. Here we ensure that a set S is a dominating set in G if and only if S has no collision in X. Then, we make a matrix X' by stacking a polynomial number of copies of X. Finally, the instance of MDL-PA is defined as (X', t), where the threshold t is set to the MDL cost due to the number of model parameters. With this construction we are able to show that G has a dominating set of size $\log n$ if and only if the MDL cost is at most t for some set of parents. We next fill in the necessary details.

Let G be a directed graph on $n-1$ nodes labeled by $1, \ldots, n-1$. We define the *reflex* of G as the $n \times n$ matrix $R = \mathrm{ref}(G)$ whose entry at the ith row and jth column, R^i_j, equals 1 if (i, j) is an arc in G or $i = j$, else R^i_j equals 0. In words, $\mathrm{ref}(G)$ is made from G by adding a new node, n, with no incoming nor outgoing arcs, and then enforcing the graph be reflexive; see Fig. 1 for an example. This matrix has a desired property, as stated in the next key lemma.

Lemma 4. *Let G be a directed graph with nodes $1, \ldots, n-1$. Then, for any subset $S \subseteq \{1, \ldots, n-1\}$, we have*

$$S \text{ is a dominating set in } G \quad \text{if and only if} \quad S \text{ has no collision in } \mathrm{ref}(G)\,.$$

Proof. Let S be a subset of $\{1, \ldots, n-1\}$. Denote $R = \mathrm{ref}(G)$ for short.

Assume first that S is a dominating set in G. Then, if S had a collision in the matrix R, we should have an index $i < n$ such that $R^i_S = R^n_S$, since only R^n_n equals 1. Accordingly, R^i_S should be a vector of 0s. But this is impossible since S is a dominating set in G, implying that G has an arc (i, j) for some $j \in S$ and, consequently, $R^i_j = 1$ by the definition of reflex.

Assume then that S is not a dominating set in G. Now it is sufficient to show that for some $i < n$ the vector R^i_S contains only 0s. Assume the contrary, that for all $i < n$ we have a $j \in S$ such that $R^i_j = 1$. But this means that every node i of G is dominated by a node $j \in S$, a contradiction. □

Let us summarize the above four lemmas:

Lemma 5. *Let G be a directed graph with nodes $1, \ldots, n-1$. Let X be the matrix obtained by stacking r copies of the reflex of G. Then, for any subset $S \subseteq \{1, \ldots, n-1\}$, we have*

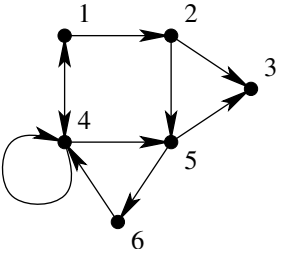
$$\begin{aligned} &\beta(\mathbf{x}_S, \mathbf{x}_n) = 0\,, \quad \text{if } S \text{ is a dominating set in } G;\\ &\beta(\mathbf{x}_S, \mathbf{x}_n) \geq 2r\,, \quad \text{if } S \text{ is not a dominating set in } G. \end{aligned}$$

Proof. Immediate from Lemmas 1, 2, 3, and 4. □

In the sequel we will use this result (Lemma 5) as a key argument. The first example of its usage is given in the proof of the next main result.

Theorem 2. MDL-PA *is LOGSNP-hard.*

(a)



(b)

$$\begin{pmatrix} 0&1&0&1&0&0 \\ 0&0&1&0&1&0 \\ 0&0&0&0&0&0 \\ 1&0&0&1&1&0 \\ 0&0&1&0&0&1 \\ 0&0&0&1&0&0 \end{pmatrix}$$

(c)

$$\begin{pmatrix} 1&1&0&1&0&0&0 \\ 0&1&1&0&1&0&0 \\ 0&0&1&0&0&0&0 \\ 1&0&0&1&1&0&0 \\ 0&0&1&0&1&1&0 \\ 0&0&0&1&0&1&0 \\ 0&0&0&0&0&0&1 \end{pmatrix}$$

Fig. 1. The reflex of a directed graph. (a) A graph with 6 nodes, (b) the adjacency matrix of the graph, and (c) the reflex of the graph. Nodes 3 and 4 form a dominating set: every other node has an arc that points to 3 or 4.

Proof. Let G be a directed graph with nodes $1, \ldots, n-1$, an instance of Log-DS. Let R be the reflex of G. Let X be the $rn \times n$ matrix obtained by stacking $r = n^2$ copies of R. Finally, set t to the value $\alpha_{\mathrm{MDL}}(rn, \log n)$.

Our first claim is that G is a positive instance of Log-DS *if* (X, t) is a positive instance of MDL-PA. Assume the latter holds. Then there exists a set of parents S such that

$$\gamma_{\mathrm{MDL}}(X, S) = \alpha_{\mathrm{MDL}}(rn, |S|) + \beta(\mathbf{x}_S, \mathbf{x}_n) \leq t = \alpha_{\mathrm{MDL}}(rn, \log n)\,.$$

Clearly, S can have at most $\log n$ elements. It remains to show that S is a dominating set in G. To see this, assume the contrary: that S is not a dominating set in G. Then, by Lemma 5, $\beta(\mathbf{x}_S, \mathbf{x}_n) \geq 2r = 2n^2$ and, thereby, $\gamma_{\mathrm{MDL}}(X, S) \geq 2n^2$. But this contradicts with the earlier conclusion that $\gamma_{\mathrm{MDL}}(X, S) \leq \alpha_{\mathrm{MDL}}(rn, \log n) \leq (1/2)\, n \log(n^3)$.

Our second claim is that G is a positive instance of Log-DS *only if* (X, t) is a positive instance of MDL-PA. Assume the former holds. Then there exists a dominating set S in G such that $|S| \leq \log n$. Now, by Lemma 5, we have $\beta(\mathbf{x}_S, \mathbf{x}_n) = 0$. Using this we see that

$$\alpha_{\mathrm{MDL}}(rn, |S|) + \beta(\mathbf{x}_S, \mathbf{x}_n) \leq t = \alpha_{\mathrm{MDL}}(rn, \log n)\,,$$

since $|S| \leq \log n$. Thus we have shown that (X, t) is a positive instance of MDL-PA.

To complete the proof, we notice that the mapping from G to X can be computed in polynomial time. Since Log-DS is LOGSNP-complete (Theorem 1), we conclude that MDL-PA is LOGSNP-hard. □

Regarding the above proof, it is worth noting that the particular choice for the number of repetitions, r, is not crucial as long as r is polynomial in n and, roughly, of order $\Omega(n \log n)$.

4 Parent Assignment Is Hard to Approximate for the MDL and AIC Costs

We next extend the result from the previous section in two dimensions. First, we show that LOGSNP-hardness holds also for other penalized maximum likelihood costs, such as AIC, that have certain properties. Second, we state the hardness result in a stronger form: the optimization variant of the parent assignment problem cannot be approximated in polynomial time to a ratio smaller than 2 unless LOGSNP = P.

We consider a generic cost function $\gamma(X, S) = \alpha(X, S) + \beta(\mathbf{x}_S, \mathbf{x}_n)$, where the penalization term $\alpha(X, S)$ is a function of the number of records m and the number of parents $k = |S|$, hence denoted as $\alpha(m, k)$. In addition, we will assume that

(A1) $\alpha(m, k)$ grows at most logarithmically in m and exponentially in k,
(A2) $\alpha(m, k)$ can be evaluated in time polynomial in m and k, and
(A3) $\alpha(m, k+1)/\alpha(m, k) \geq 2$ for all m and k.

These properties obviously hold for the MDL and AIC measures.

Proposition 1. *The functions α_{MDL} and α_{AIC} satisfy conditions (A1–A3).*

4.1 Approximating to a Ratio Less Than 2 Is Hard

We are now ready to prove the main result of this section.

Theorem 3. *Let α be a function that satisfies conditions (A1–A3). Then, for any $\epsilon > 0$, approximating* Min Parent Assignment (α) *to the ratio $2 - \epsilon$ is LOGSNP-hard.*

Proof. Assume that we have a polynomial-time algorithm $\mathcal{A}$ that, given any 0-1 input matrix X of size $n \times n$, outputs a set $S \subseteq \{1, \ldots, n-1\}$ such that

$$\gamma(X, S)/OPT(X) \leq 2 - \epsilon < 2\,,$$

for some $\epsilon > 0$; here $OPT(X)$ denotes the minimum of $\gamma(X, S')$ over all possible subsets S'.

We construct a reduction from Log-DS, similar to the one in the proof of Theorem 2. First, choose constants a and b such that $\alpha(m,k) \leq a2^{bk}\log m$ for all k and m with $k \leq m$; this we can do due to condition (A1). Then, let G be a directed graph with $n-1$ nodes, an instance of Log-DS. Let R be the reflex of G, and let X be the $rn \times n$ matrix obtained by stacking $r = n^{n+1}$ copies of R. Let S be the set given by algorithm $\mathcal{A}$ for the input X. We claim that G has a dominating set of size at most $\log n$ if and only if

$$\gamma(X,S) < 2 \cdot \alpha(rn, \log n)\,. \tag{1}$$

We prove the two directions separately. First, suppose G has a dominating set S^* of size $|S^*| \leq \log n$. Then

$$OPT(X) \leq \gamma(X,S^*) = \alpha(rn,|S^*|) \leq \alpha(rn,\log n)\,;$$

the equality follows from Lemma 5, while the last inequality is due to the monotonicity of α in the second argument (implied by (A3)). Using the approximation guarantee we obtain $\gamma(X,S) < 2 \cdot OPT(X) \leq 2 \cdot \alpha(rn,\log n)$, as desired.

For the other direction, suppose G has no dominating set of size $\log n$. Then $OPT(X) \geq \alpha(rn, 1+\log n)$, since any set S smaller than $1+\log n$ has a cost at least $\beta(\mathbf{x}_S,\mathbf{x}_n) \geq 2r = 2n^2 \geq \alpha(rn,1+\log n)$; the first inequality is by Lemma 5 and the last one is due to the choice of b (for sufficiently large n). Thus, for any set S we have that $\gamma(X,S) \geq 2 \cdot \alpha(rn,\log n)$, by condition (A3). This contradicts with inequality (1), as desired.

To complete the proof, we recall that Log-DS is LOGSNP-hard and notice that the mapping from G to X as well as the condition in inequality (1) can be computed in polynomial time. □

We notice that the main theorem of the previous section, Theorem 2, follows as a direct corollary to the above, stronger result. Let it be also noted that an even slightly stronger result holds: we may allow the number $\epsilon > 0$ in the statement of Theorem 3 depend on the instance of the Min Parent Assignment (α) problem.

By Theorem 3 and Proposition 1 we immediately have the following.

Corollary 1. *For the MDL and AIC costs, approximating* Min Parent Assignment *to a ratio less than* 2 *is LOGSNP-hard.*

4.2 Approximating to a Constant Ratio Looks Hard

Given the above hardness result, it is natural to ask whether Min Parent Assignment (α) can be approximated to any constant ratio. As we show next, the answer is likely to be negative. Namely, the positive answer would imply a polynomial-time approximation scheme (PTAS) for the following optimization version of Log Dominating Set, a problem for which no polynomial-time constant-ratio approximation algorithm is known (see Cai et al. [13]).

Min Log Dominating Set (Min-Log-DS)
Input: A directed graph G with $n-1$ nodes such that G has a dominating set of size $\log n$.
Output: A minimum-cardinality dominating set of G.

The next result provides a connection between the approximation ratio of the two problems; the result concerning constant approximation ratios follows as a corollary, as made explicit below.

Theorem 4. *Let α be a function that satisfies conditions (A1–A3). Let $c > 0$ be constant and f an integer function with $f(n) = O(n^\mu)$ for some constant $\mu \in [0, 1)$. Then* Min Parent Assignment (α) *on input matrix of size $m \times n$ cannot be approximated in polynomial time to the ratio $f(mn)$ unless* Min-Log-DS *on input graph with n nodes can be approximated in polynomial time to the ratio $1 + c \cdot \log f(n)$.*

Proof. Let us first fix some constants. Choose $\mu \in [0, 1)$ such that $f(n) \leq n^\mu$, for all sufficiently large n; this we obviously can do. In addition, choose constants a and b such that $\alpha(m, k) \leq a2^{bk} \log m$ for all k and m with $k \leq m$; this we can do due to condition (A1).

Then, assume that we have a polynomial-time algorithm $\mathcal{A}$ that, given any 0-1 input matrix X of size $n \times n$, outputs a set $S \subseteq \{1, \ldots, n-1\}$ such that $\gamma(X, S)/OPT(X) \leq f(mn)$.

We now construct a reduction from the minimum dominating set problem. We fix yet another constant $q = (b + 1 + 2\mu)/(1 - \mu)$ whose role soon becomes clear. Let G be a directed graph with $n-1$ nodes such that G has a dominating set of size $\log n$. We can assume that the smallest dominating set of G, denoted by S^*, has cardinality at least $(q + 2)/c$. Namely, this restriction obviously does not change the problem complexity (up to a polynomial factor), since one can enumerate all node subsets up to a constant cardinality in polynomial time. Let R be the reflex of G. Let X be the $rn \times n$ matrix obtained by stacking $r = n^q$ copies of R. Let S be the set given by algorithm $\mathcal{A}$ for the input X. We want to show that S is approximatively minimum dominating set of G, that is, S is a dominating set and

$$|S|/|S^*| \leq 1 + c \cdot \log f(n)\,. \tag{2}$$

To see that S is, indeed, a dominating set of G we derive a relatively small upper bound for $\gamma(X, S)$, as follows. For the optimal set S^* we have

$$\gamma(X, S^*) \leq \alpha(m, \log n) \leq an^b \log m\,,$$

which together with the assumed approximation guarantee yields

$$\begin{aligned}
\gamma(X, S) &\leq f(mn) \cdot an^b \log m \\
&\leq a(rn^2)^\mu n^b \log(rn) \\
&= a(q+1)n^{\mu(q+2)+b} \log n \\
&< n^{\mu(q+2)+b+1} \\
&= n^q\,,
\end{aligned}$$

where the strict inequality holds for large enough n (as $a(q+1)\log n = o(n)$) and the last identity holds by the choice of q. This means that S must be a dominating set of G; else, by Lemma 5, we should have $\gamma(X,S) \geq 2r = 2n^q$.

It remains to show that inequality (2) holds. To this end, we first bound

$$\begin{aligned} f(mn) &\geq \gamma(X,S)/OPT(X) \\ &= \alpha(m,|S|)/\alpha(m,|S^*|) \\ &\geq 2^{|S|-|S^*|} \\ &= 2^{|S^*|(|S|/|S^*|-1)}\,, \end{aligned}$$

where the first identity holds because S and S^* are dominating sets of G, and the second inequality is due to condition (A3). Taking logs of both sides gives us, after a little rearrangement,

$$\begin{aligned} |S|/|S^*| &\leq 1 + \frac{1}{|S^*|}\log f(mn) \\ &\leq 1 + \frac{c}{q+2}\log f(n^{q+2}) \\ &\leq 1 + c\cdot\log f(n)\,, \end{aligned}$$

since we assumed that $|S^*| \geq (q+2)/c$ and that $f(n^{q+2}) \leq f(n)^{q+2}$ (that is, f does not grow too rapidly; a polynomial f suffices here).

To complete the proof we notice that the reduction mapping can be evaluated in polynomial time. □

Corollary 2. *Let α be a function that satisfies conditions (A1–A3). Then* Min Parent Assignment (α) *cannot be approximated to any constant ratio unless* Min-Log-DS *has a polynomial-time approximation scheme.*

Proof. Suppose that Min Parent Assignment (α) can be approximated to the constant ratio $\rho > 1$ in polynomial time. Let $\epsilon > 0$ be fixed. Applying Theorem 4 with $f(n) := \rho$, for all n, and $c := \epsilon/\log\rho$ gives the approximation ratio of $1+\epsilon$ for Min-Log-DS. □

Cai et al. [13] discuss the computational complexity of Min-Log-DS. They argue that no polynomial-time algorithm can even approximate Min-Log-DS to *any* constant factor. However, the needed complexity theoretic assumptions are substantially stronger than the conventional P $\neq$ NP. Despite this gap, it is reasonable to assume that no PTAS exists for Min-Log-DS, implying the inapproximability of Min Parent Assignment (α).

5 NML Parent Assignment Is NP-Complete

In this section we show that Small Parent Assignment is NP-complete for the NML cost. Recall that this formulation of the parent assignment task assumes two input numbers: an upper bound for the cost (as in Parent Assignment)

and another upper bound for the cardinality of the parent set. The latter bound will correspond to the analogous cardinality parameter of the NP-complete Dominating Set problem [14]: Given a directed graph G and a number k, does G contain a dominating set of size at most k?

We can apply the reduction scheme presented in Sect. 3. Unlike α_{MDL} and α_{AIC}, however, α_{NML} does not have any simple, data-independent expression. Therefore, we have to work a bit to show that α_{NML} grows relatively slowly in the number of parents and in the number of data records, assuming that the data set is obtained via our reduction.

Lemma 6. *Let $r \geq 1$ be an integer and X the r-stack of a 0–1 matrix of size $n \times n$. Then, for any subset $S \subseteq \{1, \ldots, n-1\}$, we have*

$$\alpha_{NML}(X, S) \leq n \log(r+1)\,.$$

Proof. Denote by $m = rn$ the number of rows in X. Write

$$\begin{aligned} 2^{\alpha_{\mathrm{NML}}(X,S)} &= \sum_{\mathbf{x}'_n \in \{0,1\}^m} 2^{-\beta(X',S)} \\ &= \sum_{\mathbf{x}'_n \in \{0,1\}^m} \prod_{u \in \{0,1\}^{|S|} : m_u > 0} \left(\frac{m'_{u0}}{m_u}\right)^{m'_{u0}} \left(\frac{m'_{u1}}{m_u}\right)^{m'_{u1}}, \end{aligned}$$

where X' denotes the matrix obtained by replacing the nth column of X by the column $\mathbf{x}'_n$, and m'_{uv} is the number of rows in X' where the attributes x_S are set to u and the attribute x_n is set to v.

We can split the summation over $\mathbf{x}'_n$ into (at most) $2^{|S|}$ separate summations, one for each value $u \in \{0,1\}^{|S|}$ (that occurs in X). Within each summation it is sufficient to sum over the sufficient statistic m'_{u0}. Thus,

$$2^{\alpha_{\mathrm{NML}}(X,S)} = \prod_{u \in \{0,1\}^{|S|} : m_u > 0} \sum_{m'_{u0}=0}^{m_u} \binom{m_u}{m'_{u0}} \left(\frac{m'_{u0}}{m_u}\right)^{m'_{u0}} \left(\frac{m'_{u1}}{m_u}\right)^{m'_{u1}}. \qquad (3)$$

Since $\binom{k}{j} z^j (1-z)^{k-j} \leq 1$ whenever $0 \leq z \leq 1$ and $0 \leq j \leq k$, we obtain

$$2^{\alpha_{\mathrm{NML}}(X,S)} \leq \prod_{u \in \{0,1\}^{|S|} : m_u > 0} (m_u + 1)\,.$$

Finally, we examine how large a value the expression on the right-hand side can take, subject to the constraints implied by the construction: $m_u = r \cdot t_u$ with $t_u \in \{0, 1, \ldots, n\}$ and $\sum_u t_u = n$. We observe that if $t_u \geq t_w + 2$, then $(rt_u + 1)(rt_w + 1) < (r(t_u - 1) + 1)(r(t_w + 1) + 1)$. Without loss of generality we may now consider the case where u takes values from the largest possible set, $\{0,1\}^{n-1}$, in which case at least one t_u must equal 0 (for $n \geq 3$) or every t_u equals 1 (for $n \leq 2$). Consequently, the product $\prod_u (rt_u + 1)$ achieves its maximum value when each t_u is either 0 or 1. Hence,

$$2^{\alpha_{\mathrm{NML}}(X,S)} \leq (r+1)^n\,.$$

Taking logarithms on both sides gives the claimed inequality. □

This upper bound proved above is rather tight and, actually, significantly larger bounds for $\alpha(X, S)$ would already suffice for rendering Small Parent Assignment (α) NP-hard. Motivated by this fact, we formulate the hardness result in a relatively general terms.

Theorem 5. *Let $g(n) = O(poly(n))$ and $\alpha(X, S) < 2 \cdot g(n)$ whenever X is the $g(n)$-stack of a 0–1 matrix of size $n \times n$ and $S \subseteq \{1, \ldots, n-1\}$. Then* Small Parent Assignment (α) *is NP-hard.*

Proof. Let (G, k) be an instance of Dominating Set, where G is a directed graph with nodes $1, \ldots, n-1$ and k is a number between 1 and $n-1$. Set $r = g(n)$ and let X denote the r-stack of the reflex of G.

It is sufficient to show that G has a dominating set of size at most k if and only if there exists a subset $S \subseteq \{1, \ldots, n-1\}$ of size at most k such that the cost $\alpha(X, S) + \beta(\mathbf{x}_S, \mathbf{x}_n)$ is less than the threshold $t := 2 \cdot g(n) = 2r$.

Suppose first that S is a dominating set of G with $|S| \leq k$. Then, by Lemma 5, we have $\beta(\mathbf{x}_S, \mathbf{x}_n) = 0$. Since we assumed that $\alpha(X, S) < 2 \cdot g(n)$, the total cost is less than t.

Then suppose that $S \subseteq \{1, \ldots, n-1\}$ is a set with at most k elements and a cost $\alpha(X, S) + \beta(\mathbf{x}_S, \mathbf{x}_n)$ less than $t = 2 \cdot g(n) = 2r$. Then, of course, $\beta(\mathbf{x}_S, \mathbf{x}_n) < 2r$, and so, by Lemma 5, the set S is a dominating set of G. □

Now it is easy to prove the main result of this section:

Theorem 6. Small Parent Assignment (α_{NML}) *is NP-complete.*

Proof. To see NP-hardness, we use the substitution $r = g(n) = n^2$ in Lemma 6 and Theorem 5. Note that then $n \log(r+1) < 2 \cdot g(n)$.

To see that Small Parent Assignment (α_{NML}) is in NP, it is sufficient to notice that $\alpha_{\mathrm{NML}}(X, S)$, for arbitrary X and S, can be evaluated in polynomial time with respect to the size of the matrix X, for example, by using the factorization (3) in the general case of $r = 1$. □

6 Concluding Remarks

We showed that the parent assignment problem is computationally hard for some widely-used cost functions. According to the presented results, it is unlikely that one even finds a polynomial-time algorithm with a good approximation guarantee. Our reduction from the LOGSNP-hard log dominating set problem proved a relatively direct link between the two problems, however, we do not know whether the parent assignment problem for the MDL or AIC cost is LOGSNP-complete; we leave the precise complexity characterization for future research.

Our hardness results arise from three ingredients, each representing a restriction to the general parent assignment problem. Below we discuss each restriction in turn.

First, we assumed that the conditional probability model is the full-multinomial model. While this model has arguably been the most common choice in both theoretical and practical works on Bayesian networks, several other models have also

been proposed, not excluding models for continuous data. To what extend similar hardness results can be proved for those models is an open question.

Second, we considered penalized maximum-likelihood costs, such as MDL, AIC, and NML, which separate the model complexity cost and the goodness of fit in a simple manner. Other important cost functions include the Bayesian cost, which is obtained by integrating the model parameters out [1, 2]. Characterizing the complexity of parent assignment for the Bayesian cost is a natural direction for future research. Although we cannot use the key lemma (Lemma 5) as such, similar argumentation based on a reduction from the (log) dominating set problem might work. Like the NML cost, the Bayesian cost does not imply the $O(\log m)$ bound for the size of the parent set [7], which probably renders the problem NP-hard.

Third, our reduction from the dominating set problem yields hard instances that, however, do not necessary represent typical datasets one encounters in practice. This motivates seeking of appropriate constraints that would allow efficient parent assignment; works on a related large-sample setting have produced interesting characterizations of the needed assumptions and the type of optimality one can achieve [15].

Finally, it should be noted that the parent assignment problem studied in this paper falls in the broad framework of combinatorial feature selection problems (e.g., [16, 17]). Koller and Sahami [16] and Charikar et al. [17] provide insightful results concerning some interesting problem classes. However, neither of these works provides any hardness or (in)approximability result for the parent assignment problem. For *linear* classifiers (hyperplanes, perceptrons) Amaldi and Kann [18] show that finding, or approximating the number of, the relevant attributes is hard, proving that "black-box" feature selection can be hard; this result, of course, does not imply that feature selection is hard for richer hypothesis classes, e.g., the full-multinomial model.

Acknowledgements

I wish to thank David Maxwell Chickering and Christopher Meek for discussions about a large-sample variant of the parent assignment problem, which partially inspired this work. I would also like to thank an anonymous reviewer for pointing out a minor flaw in an earlier version of the proof of Theorem 3.

References

1. Cooper, G.F., Herskovits, E.: A Bayesian method for the induction of probabilistic networks from data. Machine Learning **9** (1992) 309–347
2. Heckerman, D., Geiger, D., Chickering, D.M.: Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning **20** (1995) 197–243
3. Suzuki, J.: Learning Bayesian belief networks based on the Minimun Description Length principle: An efficient algorithm using the b & b technique. In: Proceedings of the Thirteenth International Conference on Machine Learning (ICML). (1996) 462–470

4. Tian, J.: A branch-and-bound algorithm for MDL learning Bayesian networks. In: Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI), Morgan Kaufmann (2000) 580–588
5. Rissanen, J.: Modeling by shortest data description. Automatica **14** (1978) 465–471
6. Bouckaert, R.R.: Probabilistic network construction using the minimum description length principle. In: Proceedings of the European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty (ECSQARU), Springer-Verlag (1993) 41–48
7. Bouckaert, R.R.: Properties of Bayesian belief network learning algorithms. In de Mantaras, R.L., Poole, D., eds.: Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence (UAI), Morgan Kaufmann (1994) 102–109
8. Papadimitriou, C., Yannakakis, M.: On limited nondeterminism and the complexity of the V-C dimension. Journal of Computer and System Sciences **53** (1996) 161–170
9. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control **19** (1974) 716–722
10. Shtarkov, Y.M.: Universal sequential coding of single messages. Problems of Information Transmission **23** (1987) 3–17
11. Kontkanen, P., Buntine, W., Myllymäki, P., Rissanen, J., Tirri, H.: Efficient computation of stochastic complexity. In Bishop, C.M., Frey, B.J., eds.: Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics (AISTAT), Key West, FL (2003) 181–188
12. Schwarz, G.: Estimating the dimension of a model. Annals of Statistics **6** (1978) 461–464
13. Cai, L., Juedes, D., Kanj, I.: The inapproximability of non-NP-hard optimization problems. Theoretical Computer Science **289** (2002) 553–571
14. Garey, M., Johnson, D.: Computers and Intractability - A Guide to the Theory of NP-completeness. W. H. Freeman & Co., San Fransisco, CA (1971)
15. Chickering, D.M., Meek, C.: Finding optimal Bayesian networks. In: Proceedings of Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI). Morgan Kaufmann, Edmonton (2002) 94–102
16. Koller, D., Sahami, M.: Toward optimal feature selection. In: Proceedings of the Thirteenth International Conference on Machine Learning (ICML), Morgan Kaufmann (1996) 284–292
17. Charikar, M., Guruswami, V., Kumar, R., Rajagopalan, S., Sahai, A.: Combinatorial feature selection problems. In: Proceedings of the 41st IEEE Symposium on Foundations of Computer Science (FOCS), IEEE (2000) 631–640
18. Amaldi, E., Kann, V.: On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. Theoretical Computer Science **209** (1998) 237–260