When Discriminative Learning of Bayesian Network Parameters Is Easy

Hannes Wettig*, Peter Grünwald°, Teemu Roos*, Petri Myllymäki*, and Henry Tirri*

* **Complex Systems Computation Group (CoSCo)**
Helsinki Institute for Information Technology (HIIT)
University of Helsinki & Helsinki University of Technology
P.O. Box 9800, FIN-02015 HUT, Finland
{*Firstname*}.{*Lastname*}@*hiit.fi*

° **Centrum voor Wiskunde en Informatica (CWI)**
P.O. Box 94079
NL-1090 GB Amsterdam, The Netherlands.
Peter.Grunwald@cwi.nl

Abstract

Bayesian network models are widely used for discriminative prediction tasks such as classification. Usually their parameters are determined using 'unsupervised' methods such as maximization of the joint likelihood. The reason is often that it is unclear how to find the parameters maximizing the conditional (supervised) likelihood. We show how the discriminative learning problem can be solved efficiently for a large class of Bayesian network models, including the Naive Bayes (NB) and tree-augmented Naive Bayes (TAN) models. We do this by showing that under a certain general condition on the network structure, the discriminative learning problem is exactly equivalent to logistic regression with unconstrained convex parameter spaces. Hitherto this was known only for Naive Bayes models. Since logistic regression models have a concave log-likelihood surface, the global maximum can be easily found by local optimization methods.

1 Introduction

In recent years it has been recognized that for discriminative prediction tasks such as classification, we should use a 'supervised' learning algorithm such as conditional likelihood maximization [Friedman *et al.*, 1997; Ng and Jordan, 2001; Kontkanen *et al.*, 2001; Greiner and Zhou, 2002]. Nevertheless, for Bayesian network models the parameters are customarily determined using ordinary methods such as maximization of the joint (unsupervised) likelihood. One of the main reasons for this discrepancy is the difficulty in finding the global maximum of the conditional likelihood. In this paper, we show that this problem can be solved, as long as the underlying Bayesian network meets a particular additional condition, which is satisfied for many existing Bayesian-network based models including Naive Bayes (NB), TAN (tree-augmented NB) and 'diagnostic' models [Kontkanen *et al.*, 2001].

We consider domains of discrete-valued random variables. We find the maximum conditional likelihood parameters by logarithmic reparametrization. In this way, each conditional Bayesian network model is mapped to a logistic regression model, for which the likelihood surface is known to be concave. However, in some cases the parameters of this logistic regression model are not allowed to vary freely. In other words, the Bayesian network model corresponds to a subset of a logistic regression model rather than the full model.

Our main result (Thm. 3 below) provides a general condition on the network structure under which, as we prove, the Bayesian network model is mapped to a full logistic regression model with freely varying parameters. Therefore, in the new parametrization the conditional log-likelihood becomes a concave function of the parameters that under our condition are allowed to vary freely over the convex set $\mathbb{R}^k$. Now we can find the global maximum in the conditional likelihood surface by simple local optimization techniques such as hill climbing.

The result still leaves open the possibility that there are *no* network structures for which the conditional likelihood surface has local, non-global maxima. This would make our condition superfluous. Our second result (Thm. 4 below) shows that this is not the case: there are very simple network structures that do not satisfy our condition, and for which the conditional likelihood can exhibit local, non-global maxima.

Viewing Bayesian network (BN) models as subsets of logistic regression models is not new; it was done earlier in papers such as [Heckerman and Meek, 1997a; Ng and Jordan, 2001; Greiner and Zhou, 2002]. Also, the concavity of the log-likelihood surface for logistic regression is known. Our main contribution is to supply the condition under which Bayesian network models correspond to logistic regression with *completely freely varying parameters*. Only then can we guarantee that there are no local maxima in the likelihood surface. As a direct consequence of our result, we show for the first time that the supervised likelihood of, for instance, the tree-augmented Naive Bayes (TAN) model has no local maxima.

This paper is organized as follows. In Section 2 we introduce Bayesian networks and an alternative, so-called L-parametrization. In Section 3 we show that this allows us to consider Bayesian network models as logistic regression models. Based on earlier results in logistic regression, we conclude that in the L-parametrization the supervised log-likelihood is a concave function. In Section 4 we present our main results, giving conditions under which the two parametrizations correspond to exactly the same conditional distributions. Conclusions are summarized in Section 5; proofs of the main results are given in Appendix A.

2 Bayesian Networks and the L-model

We assume that the reader is familiar with the basics of the theory of Bayesian networks, see, e.g., [Pearl, 1988].

Consider a random vector $X = (X_0, X_1, \ldots, X_{M'})$, where each variable X_i takes values in $\{1, \ldots, n_i\}$. Let $\mathcal{B}$ be a Bayesian network structure over X, that factorizes $P(X)$ into

$$P(X) = \prod_{i=0}^{M'} P(X_i \mid Pa_i), \tag{1}$$

where $Pa_i \subseteq \{X_0, \ldots, X_{M'}\}$ is the parent set of variable X_i in $\mathcal{B}$.

We are interested in predicting some *class variable* X_m for some $m \in \{0, \ldots, M'\}$ conditioned on all X_i, $i \neq m$. Without loss of generality we may assume that $m = 0$ (i.e., X_0 is the class variable) and that the children of X_0 in $\mathcal{B}$ are $\{X_1, \ldots, X_M\}$ for some $M \leq M'$. For instance, in the so-called Naive Bayes model we have $M = M'$ and the children of the class variable X_0 are independent given the value of X_0. The Bayesian network model corresponding to $\mathcal{B}$ is the set of all distributions satisfying the conditional independencies encoded in $\mathcal{B}$. It is usually parametrized by vectors $\Theta^{\mathcal{B}}$ with components of the form $\theta^{\mathcal{B}}_{x_i|pa_i}$ defined by

$$\theta^{\mathcal{B}}_{x_i|pa_i} := P(X_i = x_i \mid Pa_i = pa_i), \tag{2}$$

where pa_i is any configuration (set of values) for the parents Pa_i of X_i. Whenever we want to emphasize that each pa_i is determined by the complete data vector $x = (x_0, \ldots, x_{M'})$, we write $pa_i(x)$ to denote the configuration of Pa_i in $\mathcal{B}$ given by the vector x. For a given data vector $x = (x_0, x_1, \ldots, x_{M'})$, we sometimes need to consider a modified vector where x_0 is replaced by x'_0 and the other entries remain the same. We then write $pa_i(x'_0, x)$ for the configuration of Pa_i given by $(x'_0, x_1, \ldots, x_{M'})$.

We let $\mathcal{M}^{\mathcal{B}}$ be the set of *conditional* distributions $P(X_0 \mid X_1, \ldots, X_{M'}, \Theta^{\mathcal{B}})$ corresponding to distributions $P(X_0, \ldots, X_{M'} \mid \Theta^{\mathcal{B}})$ satisfying the conditional independencies encoded in $\mathcal{B}$. The conditional distributions in $\mathcal{M}^{\mathcal{B}}$ can be written as

$$P(x_0 \mid x_1, \ldots, x_{M'}, \Theta^{\mathcal{B}}) = \frac{\theta^{\mathcal{B}}_{x_0|pa_0(x)} \prod_{i=1}^{M'} \theta^{\mathcal{B}}_{x_i|pa_i(x)}}{\sum_{x'_0=1}^{n_0} \theta^{\mathcal{B}}_{x'_0|pa_0(x)} \prod_{i=1}^{M'} \theta^{\mathcal{B}}_{x_i|pa_i(x'_0,x)}}, \tag{3}$$

extended to N outcomes by independence.

Given a complete data-matrix $D = (x^1, \ldots, x^N)$, the *conditional log-likelihood*, $S^{\mathcal{B}}(D; \Theta^{\mathcal{B}})$, with parameters $\Theta^{\mathcal{B}}$ is given by

$$S^{\mathcal{B}}(D; \Theta^{\mathcal{B}}) := \sum_{j=1}^{N} S^{\mathcal{B}}(x^j; \Theta^{\mathcal{B}}), \tag{4}$$

where

$$S^{\mathcal{B}}(x; \Theta^{\mathcal{B}}) := \ln P(x_0 \mid x_1, \ldots, x_{M'}, \Theta^{\mathcal{B}}). \tag{5}$$

Note that in (3), and hence also in (4), all $\theta^{\mathcal{B}}_{x_i|pa_i}$ with $i > M$ (standing for nodes that are neither the class variable nor any of its children) cancel out, since for these terms we have $pa_i(x) \equiv pa_i(x'_0, x)$ for all x'_0. Thus the only relevant parameters for determining the conditional likelihood are of the form $\theta^{\mathcal{B}}_{x_i|pa_i}$ with $i \in \{0, \ldots, M\}$, $x_i \in \{1, \ldots, n_i\}$ and pa_i any configuration of parents Pa_i. We order these parameters lexicographically and define $\mathbf{\Theta}^{\mathcal{B}}$ to be the set of vectors constructed this way, with $\theta^{\mathcal{B}}_{x_i|pa_i} > 0$ and $\sum_{x_i=1}^{n_i} \theta^{\mathcal{B}}_{x_i|pa_i} = 1$ for all $i \in \{0, \ldots, M\}$, x_i and all values (configurations) of pa_i. Note that we require all parameters to be *strictly* positive.

The model $\mathcal{M}^{\mathcal{B}}$ does not contain any notion of the joint distribution: Terms such as $P(X_i \mid Pa_i)$, where $0 < i \leq M'$, are undefined and neither are we interested in them. Our task is prediction of X_0 *given* $X_1, \ldots, X_{M'}$. Heckerman and Meek call such models *Bayesian regression/classification* (BRC) models [Heckerman and Meek, 1997a; 1997b].

For an arbitrary conditional Bayesian network model $\mathcal{M}^{\mathcal{B}}$, we now define the so-called L-model, another set of conditional distributions $P(X_0 \mid X_1, \ldots, X_{M'})$. This model, which we denote by $\mathcal{M}^L$, is parametrized by vectors Θ^L in some set $\mathbf{\Theta^L}$ that closely resembles $\mathbf{\Theta}^{\mathcal{B}}$. Each different $\mathcal{M}^{\mathcal{B}}$ gives rise to a corresponding $\mathcal{M}^L$, although we do not necessarily have $\mathcal{M}^{\mathcal{B}} = \mathcal{M}^L$. For each component $\theta^{\mathcal{B}}_{x_i|pa_i}$ of each vector $\Theta^{\mathcal{B}} \in \mathbf{\Theta}^{\mathcal{B}}$, there is a corresponding component $\theta^L_{x_i|pa_i}$ of the vectors $\Theta^L \in \mathbf{\Theta^L}$. The components $\theta^L_{x_i|pa_i}$ take values in the range $(-\infty, \infty)$ rather than $(0, 1)$. Each vector $\Theta^L \in \mathbf{\Theta^L}$ defines the following conditional distribution:

$$P(x_0 \mid x_1, \ldots, x_{M'}, \Theta^L) := \frac{(\exp \theta^L_{x_0|pa_0(x)}) \prod_{i=1}^{M} \exp \theta^L_{x_i|pa_i(x)}}{\sum_{x'_0=1}^{n_0} (\exp \theta^L_{x'_0|pa_0(x)}) \prod_{i=1}^{M} \exp \theta^L_{x_i|pa_i(x'_0,x)}}. \tag{6}$$

The model $\mathcal{M}^L$ is the set of conditional distributions $P(X_0 \mid X_1, \ldots, X_{M'}, \Theta^L)$ indexed by $\Theta^L \in \mathbf{\Theta^L}$, extended to N outcomes by independence. Given a data-matrix D, let $S^L(D; \Theta^L)$ be the conditional log-likelihood with parameters Θ^L, defined analogously to (4) with (6) in place of (3).

Theorem 1. $\mathcal{M}^{\mathcal{B}} \subseteq \mathcal{M}^L$.

Proof. From (3) and (6) we get that Θ^L defined by setting $\theta^L_{x_i|pa_i} = \ln \theta^{\mathcal{B}}_{x_i|pa_i}$ for all i, x_i and pa_i, indexes the same conditional distribution as $\Theta^{\mathcal{B}}$. ∎

In words, all the conditional distributions that can be represented by parameters $\Theta^{\mathcal{B}} \in \mathbf{\Theta}^{\mathcal{B}}$ can also be represented by parameters $\Theta^L \in \mathbf{\Theta^L}$. The converse of Theorem 1, i.e., $\mathcal{M}^L \subseteq \mathcal{M}^{\mathcal{B}}$, is true only under some additional conditions on the network structure, as we explain in Section 4. First we take a closer look at the L-model.

3 The L-model Viewed as Logistic Regression

Although L-models are closely related to and in some cases formally identical to Bayesian network models, we can also think of them as predictors that combine the information of the attributes using the so-called *softmax* rule [Heckerman and Meek, 1997b; Ng and Jordan, 2001]. In statistics, such models have been extensively studied under the name of *logistic regression* models, see, e.g. [McLachlan, 1992, p.255].

More precisely, let $X_0 = \{1, \dots, n_0\}$ and let $Y_1, \dots, Y_k$ be real-valued random variables. The *multiple logistic regression model with dependent variable X_0 and covariates* $Y_1, \dots, Y_k$ is defined as the set of conditional distributions

$$P(x_0 \mid y_1, \dots, y_k) := \frac{\exp \sum_{i=1}^{k} \beta_{x_0|i}\, y_i}{\sum_{x'_0=1}^{n_0} \exp \sum_{i=1}^{k} \beta_{x'_0|i}\, y_i} \qquad (7)$$

where the $\beta_{x_0|i}$ are allowed to take on all values in $\mathbb{R}$. This defines a conditional model parameterized in $\mathbb{R}^{n_0 \cdot k}$. Now, for $i \in \{0, \dots, M\}$, $x_i \in \{1, \dots, n_i\}$ and pa_i in the set of parent configurations of X_i, let

$$Y_{(x_i, pa_i)} := \begin{cases} 1 & \text{if } X_i = x_i \text{ and } Pa_i = pa_i \\ 0 & \text{otherwise.} \end{cases} \qquad (8)$$

The indicator random variables $Y_{(x_i, pa_i)}$ thus obtained can be lexicographically ordered and renamed $1, \dots, k$, which shows that each L-model corresponding to a Bayesian network structure $\mathcal{B}$ as in (6) is formally identical to the logistic model (7) with dependent variable X_0 and covariates given by (8). So, for all network structures $\mathcal{B}$, the corresponding L-model $\mathcal{M}^L$ is the standard multiple logistic model, where the input variables for the logistic model are transformations of the input variables to the L-model, the transformation being determined by the network structure $\mathcal{B}$.

It turns out that the conditional log-likelihood in the L-parametrization is a concave function of the parameters:

Theorem 2. *The parameter set $\mathbf{\Theta^L}$ is convex, and the conditional log-likelihood $S^L(D; \Theta^L)$ is concave, though not strictly concave.*

Proof. The first part is obvious since each parameter can take values in $(-\infty, \infty)$. Concavity of $S^L(D; \Theta^L)$ is a direct consequence of the fact that multiple logistic regression models are exponential families; see, e.g., [McLachlan, 1992, p.260]. For an example showing that the conditional log-likelihood is not strictly concave, see [Wettig *et al.*, 2002]. ∎

Remark. Non-strictness of the proven concavity may pose a technical problem in optimization. This can be avoided by assigning a strictly concave prior on the model parameters and then maximizing the 'conditional posterior' [Grünwald *et al.*, 2002; Wettig *et al.*, 2002] instead of the likelihood. One may also prune the model of weakly supported parameters and/or add constraints to arrive at a strictly concave conditional likelihood surface. Our experiments [Wettig *et al.*, 2002] suggest that for small data samples this should be done in any case, in order to avoid over-fitting; see also Section 5. Any constraint added should of course leave the parameter space a convex set, e.g. a subspace of the full $\mathbf{\Theta^L}$.

Corollary 1. *There are no local, non-global, maxima in the likelihood surface of an L-model.*

The conditions under which a global maximum exists are discussed in, e.g., [McLachlan, 1992] and references therein. A possible solution in cases where no maximum exists is to introduce a strictly concave prior as discusssed above.

The global conditional maximum likelihood parameters obtained from training data can be used for prediction of future data. In addition, as discussed in [Heckerman and Meek, 1997a], they can be used to perform model selection among several competing model structures using, e.g., the BIC or (approximate) MDL criteria. In [Heckerman and Meek, 1997a] it is stated that for general conditional Bayesian network models $\mathcal{M}^{\mathcal{B}}$, "although it may be difficult to determine a global maximum, gradient-based methods [...] can be used to locate local maxima". Theorem 2 shows that if the network structure $\mathcal{B}$ is such that the two models are equivalent, $\mathcal{M}^{\mathcal{B}} = \mathcal{M}^L$, we can find even the *global* maximum of the conditional likelihood by reparametrizing $\mathcal{M}^{\mathcal{B}}$ to the L-model, and using some local optimization method. Thus, the question under which condition $\mathcal{M}^{\mathcal{B}} = \mathcal{M}^L$ becomes crucial. It is this question we address in the next section.

Remark. Because the log-transformation is continuous, it follows (with some calculus) that, if $\mathcal{M}^{\mathcal{B}} = \mathcal{M}^L$, then all maxima of the (concave) conditional likelihood in the L-parameterization are global (and connected) maxima also in the original parametrization. Nevertheless, the likelihood surface as a function of $\Theta^{\mathcal{B}} \in \mathbf{\Theta}^{\mathcal{B}}$ has some unpleasant properties (see [Wettig *et al.*, 2002]): it is *not* concave in general and, worse, it can have 'wrinkles': by these we mean convex subsets $\mathbf{\Theta}_0^{\mathcal{B}}$ of $\mathbf{\Theta}^{\mathcal{B}}$, such that, under the constraint that $\Theta^{\mathcal{B}} \in \mathbf{\Theta}_0^{\mathcal{B}}$, the likelihood surface does exhibit local, non-global maxima. This suggests that it is computationally preferrable to optimize over $\mathbf{\Theta^L}$ rather than $\mathbf{\Theta}^{\mathcal{B}}$. Empirical evidence for this is reported in [Greiner and Zhou, 2002].

4 Main Result

By setting $\theta^L_{x_i|pa_i} = \ln \theta^{\mathcal{B}}_{x_i|pa_i}$, it follows that each distribution in $\mathcal{M}^{\mathcal{B}}$ is also in $\mathcal{M}^L$ (Thm. 1). This suggests that, by doing the reverse transformation

$$\theta^{\mathcal{B}}_{x_i|pa_i} = \exp \theta^L_{x_i|pa_i}, \qquad (9)$$

we could also show that distributions in $\mathcal{M}^L$ are also in $\mathcal{M}^{\mathcal{B}}$. However, $\mathbf{\Theta^L}$ contains distributions that violate the 'sum-up-to-one constraint', i.e., for some $\Theta^L \in \mathbf{\Theta^L}$ we have $\sum_{x_i=1}^{n_i} \exp \theta^L_{x_i|pa_i} \neq 1$ for some $i \in \{0, \dots, M'\}$ and pa_i. Then the corresponding $\Theta^{\mathcal{B}}$ is *not* in $\mathbf{\Theta}^{\mathcal{B}}$. But, since the L-parameterization is redundant (many different Θ^L index the same conditional distribution $P(\cdot \mid \cdot) \in \mathcal{M}^L$), it may still be the case that the *distribution* $P(\cdot \mid \cdot, \Theta^L)$ *indexed by* Θ^L is in $\mathcal{M}^{\mathcal{B}}$. Indeed, it turns out that for *some* network structures $\mathcal{B}$, the corresponding $\mathcal{M}^L$ is such that *each* distribution in $\mathcal{M}^L$ can be expressed by a parameter vector Θ^L such that $\sum_{x_i=1}^{n_i} \exp \theta^L_{x_i|pa_i} = 1$ for all $i \in \{0, \dots, M'\}$ and pa_i. In that case, by (9), we *do* have $\mathcal{M}^{\mathcal{B}} = \mathcal{M}^L$. Our main result is that this is the case if $\mathcal{B}$ satisfies the following condition:

Condition 1. For all $j \in \{1, \dots, M\}$, there exists $X_i \in Pa_j$ such that $Pa_j \subseteq Pa_i \cup \{X_i\}$.

Remark. Condition 1 implies that the class X_0 must be a 'moral node', i.e., it cannot have a common child with a node it is not directly connected with. But Condition 1 demands more than that; see Figures 1 and 2.

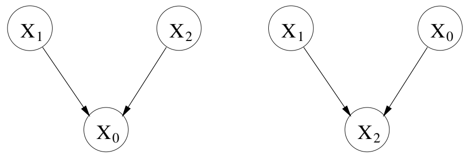

Figure 1: A simple Bayesian network (the class variable is denoted by X_0) satisfying Condition 1 (left); and a network that does not satisfy the condition (right).

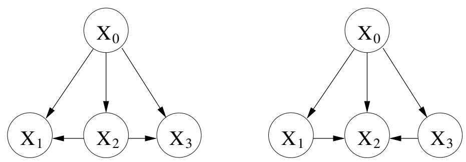

Figure 2: A tree-augmented Naive Bayes (TAN) model satisfying Condition 1 (left); and a network that is not TAN (right). Even though in both cases the class variable X_0 is a moral node, the network on the right does not satisfy Condition 1.

Example 1. Consider the Bayesian networks depicted in Figure 1. The leftmost network, $\mathcal{B}_1$, satisfies Condition 1, the rightmost network, $\mathcal{B}_2$, does not. Theorem 4 shows that the conditional likelihood surface of $\mathcal{M}^{\mathcal{B}_2}$ can have local maxima, implying that in this case $\mathcal{M}^{\mathcal{B}} \neq \mathcal{M}^L$. $\diamond$

Examples of network structures that satisfy Condition 1 are the Naive Bayes (NB) and the tree-augmented Naive Bayes (TAN) models [Friedman *et al.*, 1997]. The latter is a generalization of the former in which the children of the class variable are allowed to form tree-structures; see Figure 2.

Proposition 1. *Condition 1 is satisfied by the Naive Bayes and the tree-augmented Naive Bayes structures.*

Proof. For Naive Bayes, we have $Pa_j \subseteq \{X_0\}$ for all $j \in \{1, \ldots, M\}$. For TAN models, all children of the class variable have either one or two parents. For children with only one parent (the class variable) we can use the same argument as in the NB case. For any child X_j with two parents, let X_i be the parent that is not the class variable. Because X_i is also a child of the class variable, we have $Pa_j \subseteq Pa_i \cup \{X_i\}$. ∎

Condition 1 is also automatically satisfied if X_0 only has incoming arcs[1] ('diagnostic' models, see [Kontkanen *et al.*, 2001]). For Bayesian network structures $\mathcal{B}$ for which the condition does not hold, we can always add some arcs to arrive at a structure $\mathcal{B}'$ for which the condition does hold (for instance, add an arc from X_1 to X_3 in the rightmost network in Figure 2). Therefore, $\mathcal{M}^{\mathcal{B}}$ is always a subset of a larger model $\mathcal{M}^{\mathcal{B}'}$ for which the condition holds. We are now ready to present our main result (for proof see Appendix A):

Theorem 3. *If $\mathcal{B}$ satisfies Condition 1, then $\mathcal{M}^{\mathcal{B}} = \mathcal{M}^L$.*

Together with Corollary 1, Theorem 3 shows that Condition 1 suffices to ensure that the conditional likelihood surface of $\mathcal{M}^{\mathcal{B}}$ has no local (non-global) maxima. Proposition 1 now implies that, for example, the conditional likelihood surface of TAN models has no local maxima. Therefore, a global maximum can be found by local optimization techniques.

But what about the case in which Condition 1 does not hold? Our second result, Theorem 4 (proven in Appendix A) says that in this case, there can be local maxima:

Theorem 4. *Let $\mathcal{B}_2 = X_1 \rightarrow X_2 \leftarrow X_0$ be the network structure depicted in Figure 1 (right). There exist data samples such that the conditional likelihood has local, non-global maxima over $\mathcal{M}^{\mathcal{B}_2}$.*

The theorem implies that $\mathcal{M}^L \neq \mathcal{M}^{\mathcal{B}_2}$. Thus, Condition 1 is not superfluous. We may now ask whether our condition is *necessary* for having $\mathcal{M}^L = \mathcal{M}^{\mathcal{B}}$; that is, whether $\mathcal{M}^L \neq \mathcal{M}^{\mathcal{B}}$ for *all* network structures that violate the condition. We plan to address this intriguing open question in future work.

5 Concluding Remarks

We showed that one can effectively find the parameters maximizing the conditional (supervised) likelihood of NB, TAN and many other Bayesian network models. We did this by showing that the network structure of these models satisfies our 'Condition 1', which ensures that the conditional distributions corresponding to such models are equivalent to a particular multiple logistic regression model with unconstrained parameters. An arbitrary network structure can always be made to satisfy Condition 1 by adding arcs. Thus, we can embed any Bayesian network model in a larger model (with less independence assumptions) that satisfies Condition 1.

Test runs for the Naive Bayes case in [Wettig *et al.*, 2002] have shown that maximizing the conditional likelihood in contrast to the usual practice of maximizing the joint (unsupervised) likelihood is feasible and yields greatly improved classification. Similar results are reported in [Greiner and Zhou, 2002]. Our conclusions are also supported by theoretical analysis in [Ng and Jordan, 2001]. Only on very small data sets we sometimes see that joint likelihood optimization outperforms conditional likelihood, the reason apparently being that the conditional method is more inclined to overfitting. We conjecture that in such cases, rather than resorting to maximizing the joint instead of the conditional likelihood, it may be preferable to use a simpler model or simplify (i.e. prune or restrict) the model at hand and still choose its parameters in a discriminative fashion. In our setting, this would amount to model selection using the L-parametrization. This is a subject of our future research.

References

[Friedman *et al.*, 1997] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.

[Greiner and Zhou, 2002] R. Greiner and W. Zhou. Structural extension to logistic regression: Discriminant parameter learning of belief net classifiers. In *Proceedings of the Eighteenth Annual National Conference on Artificial Intelligence (AAAI-02)*, pages 167–173, Edmonton, 2002.

[Grünwald *et al.*, 2002] P. Grünwald, P. Kontkanen, P. Myllymäki, T. Roos, H. Tirri, and H. Wettig. Supervised posterior distri-

[1] It is easy to see that in that case the maximum conditional likelihood parameters may even be determined analytically.

butions, 2002. Presented at the Seventh Valencia International Meeting on Bayesian Statistics, Tenerife, Spain.

[Heckerman and Meek, 1997a] D. Heckerman and C. Meek. Embedded bayesian network classifiers. Technical Report MSR-TR-97-06, Microsoft Research, 1997.

[Heckerman and Meek, 1997b] D. Heckerman and C. Meek. Models and selection criteria for regression and classification. In D. Geiger and P. Shenoy, editors, *Uncertainty in Arificial Intelligence 13*, pages 223–228. Morgan Kaufmann Publishers, San Mateo, CA, 1997.

[Kontkanen *et al.*, 2001] P. Kontkanen, P. Myllymäki, and H. Tirri. Classifier learning with supervised marginal likelihood. In J. Breese and D. Koller, editors, *Proceedings of the 17th International Conference on Uncertainty in Artificial Intelligence (UAI'01)*. Morgan Kaufmann Publishers, 2001.

[McLachlan, 1992] G.J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, New York, 1992.

[Ng and Jordan, 2001] A.Y. Ng and M.I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *Advances in Neural Information Processing Systems*, 14:605–610, 2001.

[Pearl, 1988] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Mateo, CA, 1988.

[Wettig *et al.*, 2002] H. Wettig, P. Grünwald, T. Roos, P. Myllymäki, and H. Tirri. On supervised learning of Bayesian network parameters. Technical Report HIIT-2002–1, Helsinki Institute for Information Technology (HIIT), 2002. Available at http://cosco.hiit.fi/Articles/hiit2002-1.ps.

A Proofs

Proof of Theorem 3. We introduce some more notation. For $j \in \{1,\ldots,M\}$, let m_j be the maximum number in $\{0,\ldots,M\}$ such that $X_{m_j} \in Pa_j$, $Pa_j \subseteq Pa_{m_j} \cup \{X_{m_j}\}$. Such an m_j exists by Condition 1. To see this, note that the $X_i \in Pa_j$ mentioned in Condition 1 must lie in the set $\{X_0, X_1, \ldots, X_m\}$ (otherwise $X_0 \in Pa_j, X_0 \notin Pa_i$, so $Pa_j \not\subseteq Pa_i \cup \{X_i\}$, contradiction).

Condition 1 implies that pa_j is completely determined by the pair (x_{m_j}, pa_{m_j}). We can therefore introduce functions Q_j mapping (x_{m_j}, pa_{m_j}) to the corresponding pa_j. Hence, for all $x = (x_0, \ldots, x_{M'})$ and $j \in \{1,\ldots,M\}$ we have

$$pa_j = Q_j(x_{m_j}, pa_{m_j}). \tag{10}$$

We introduce, for all $i \in \{0,\ldots,M\}$ and for each configuration pa_i of Pa_i, a constant $c_{i|pa_i}$ and define, for any $\Theta^L \in \mathbf{\Theta^L}$,

$$\theta^{(c)}_{x_i|pa_i} := \theta^L_{x_i|pa_i} + c_{i|pa_i} - \sum_{j:m_j=i} c_{j|Q_j(x_i,pa_i)}. \tag{11}$$

The parameters $\theta^{(c)}_{x_i|pa_i}$ constructed this way are combined to a vector $\Theta^{(c)}$ which is clearly a member of $\mathbf{\Theta^L}$.

Having introduced this notation, we now show that no matter how we choose the constants $c_{i|pa_i}$, for all Θ^L and corresponding $\Theta^{(c)}$ we have $S^L(D;\Theta^{(c)}) = S^L(D;\Theta^L)$.

We first show that, for all possible vectors x and the corresponding parent configurations, no matter how the $c_{i|pa_i}$ are chosen, it holds that

$$\sum_{i=0}^{M} \theta^{(c)}_{x_i|pa_i} = \sum_{i=0}^{M} \theta^L_{x_i|pa_i} + c_{0|pa_0}. \tag{12}$$

To derive (12) we substitute all terms of $\sum_{i=0}^M \theta^{(c)}_{x_i|pa_i}$ by their definition (11). Clearly, for all $j \in \{1,\ldots,M\}$, there is exactly one term of the form $c_{j|pa_j}$ that appears in the sum with a positive sign. Since for each $j \in \{1,\ldots,M\}$ there exists exactly one $i \in \{0,\ldots,M\}$ with $m_j = i$, it must be the case that for all $j \in \{1,\ldots,M\}$, a term of the form $c_{j|Q_j(x_i,pa_i)}$ appears exactly once in the sum with a negative sign. By (10) we have $c_{j|Q_j(x_i,pa_i)} = c_{j|pa_j}$. Therefore all terms $c_{j|pa_j}$ that appear once with a positive sign also appear once with a negative sign. It follows that, except for $c_{0|pa_0}$, all terms $c_{j|pa_j}$ cancel. This establishes (12). By plugging in (12) into (6), it follows that $S^L(D;\Theta^{(c)}) = S^L(D;\Theta^L)$ for all D.

Now set, for all x_i and pa_i,

$$\theta^{\mathcal{B}}_{x_i|pa_i} = \exp \theta^{(c)}_{x_i|pa_i}. \tag{13}$$

We show that we can determine the constants $c_{i|pa_i}$ such that for all $i \in \{0,\ldots,M\}$ and pa_i, the 'sum up to one' constraint is satisfied, i.e., we have

$$\sum_{x_i=1}^{n_i} \theta^{\mathcal{B}}_{x_i|pa_i} = 1. \tag{14}$$

We achieve this by sequentially determining values for $c_{i|pa_i}$ in a particular order.

We need some additional terminology: we say 'c_i *is determined'* if for all configurations pa_i of Pa_i, we have already determined $c_{i|pa_i}$. We say 'c_i *is undetermined'* if we have determined $c_{i|pa_i}$ for *no* configuration pa_i of Pa_i. We say 'c_i is *ready to be determined'* if c_i is undetermined and at the same time all c_j with $m_j = i$ have been determined.

We note that as long as some c_i with $i \in \{0,\ldots,M\}$ are undetermined, there must exist $c_{i'}$ that are ready to be determined. To see this, first take any $i \in \{0,\ldots,M\}$ with c_i undetermined. Either c_i itself is ready to be determined (in which case we are done), or there exists $j \in \{1,\ldots M\}$ with $m_j = i$ (and hence $X_i \in Pa_j$) such that c_j is undetermined. If c_j is ready to be determined, we are done. Otherwise we repeat the argument, move forward in $\mathcal{B}$ restricted to $\{X_0,\ldots,X_M\}$ and (because $\mathcal{B}$ is acyclic) within M steps surely find a c_l that is ready to be determined.

We now describe an algorithm that sequentially assigns values to $c_{i|pa_i}$ such that (14) is satisfied. We start with all c_i undetermined and repeat the following steps:

WHILE there exists c_i, $i \in \{0,\ldots,M\}$, that is undetermined DO

1. Pick the largest i such that c_i is ready to be determined.
2. Set, for all configurations pa_i of Pa_i, $c_{i|pa_i}$ such that $\sum_{x_i=1}^{n_i} \theta^{\mathcal{B}}_{x_i|pa_i} = 1$ holds.

DONE

The algorithm loops $M+1$ times and then halts. Step 2 does not affect the values of $c_{j|pa_j}$ for any j, pa_j such that $c_{j|pa_j}$ has already been determined. Therefore, after the algorithm halts, (14) holds.

Let $\Theta^L \in \mathbf{\Theta^L}$. Each choice of constants $c_{i|pa_i}$ determines a corresponding vector $\Theta^{(c)}$ with components given by (11). This in turn determines a corresponding vector $\Theta^{\mathcal{B}}$ with components given by (13). In Stage 2 we showed that we can take the $c_{i|pa_i}$ such that (14) holds. This is the choice of $c_{i|pa_i}$ which we adopt. With this particular choice, $\Theta^{\mathcal{B}}$ indexes a distribution in $\mathcal{M}^{\mathcal{B}}$. By applying the log-transformation to the components of $\Theta^{\mathcal{B}}$ we find that for any D of any length, $S^{\mathcal{B}}(D;\Theta^{\mathcal{B}}) = S^L(D;\Theta^{(c)})$, where $S^{\mathcal{B}}(D;\Theta^{\mathcal{B}})$ denotes the conditional log-likelihood of $\Theta^{\mathcal{B}}$ as given by summing the logarithm of (3). The result of Stage 1 now implies that $\Theta^{\mathcal{B}}$ indexes the same conditional distribution as Θ^L. Since $\Theta^L \in \mathbf{\Theta^L}$ was chosen arbitrarily, this shows that $\mathcal{M}^L \subseteq \mathcal{M}^{\mathcal{B}}$. Together with Theorem 1 this concludes the proof. ∎

Proof (sketch) of Theorem 4. Use the rightmost network in Figure 1 with structure $X_0 \to X_2 \leftarrow X_1$. Let the data be $D = ((1,1,1),(1,1,2),(2,2,1),(2,2,2))$. Note that X_0 and X_1 always have the same value. We first show that with this data, there are four local, non-connected suprema of the conditional likelihood.

We are interested in predicting the value of X_0 given X_1, and X_2. The parameter defining the distribution of X_1 has no effect on conditional predictions and we can ignore it. For the remaining five parameters we use the following notation:

$$\begin{aligned}
\theta_2 &:= P(X_0 = 2), \\
\theta_{2|1,1} &:= P(X_2 = 2 \mid X_0 = 1, X_1 = 1), \\
\theta_{2|1,2} &:= P(X_2 = 2 \mid X_0 = 1, X_1 = 2), \\
\theta_{2|2,1} &:= P(X_2 = 2 \mid X_0 = 2, X_1 = 1), \\
\theta_{2|2,2} &:= P(X_2 = 2 \mid X_0 = 2, X_1 = 2).
\end{aligned} \tag{15}$$

The conditional log-likelihood can be written as

$$S^{\mathcal{B}}(D;\Theta^{\mathcal{B}}) = g(1-\theta_2, \theta_{2|1,1}, \theta_{2|2,1}) + g(\theta_2, \theta_{2|2,2}, \theta_{2|1,2}), \tag{16}$$

where

$$g(x,y,z) := f(x,y,z) + f(x, 1-y, 1-z), \tag{17}$$

and

$$f(x,y,z) := \ln \frac{xy}{xy + (1-x)z}. \tag{18}$$

Figure 3 illustrates functions $g(x,y,z)$ at $x = 0.5$. In (16) each parameter except θ_2 appears only once. Thus, for a fixed θ_2 we can maximize each term separately. ¿From Lemma 1 below it follows that the supremum of the log-likelihood with θ_2 fixed is $\ln(1-\theta_2) + \ln(\theta_2)$, which achieves its maximum value $-2\ln 2$ at $\theta_2 = 0.5$. Furthermore, the lemma shows that the log-likelihood approaches its supremum when $\theta_{2|2,1} \in \{0,1\}$, $\theta_{2|1,2} \in \{0,1\}$, $\theta_{2|1,1} \to \theta_{2|2,1}$, and $\theta_{2|2,2} \to \theta_{2|1,2}$.

Setting $y = 0.5$ results in

$$\sup_{0 \le z \le 1} g(x, 0.5, z) = \ln \frac{x}{2-x} < \ln x. \tag{19}$$

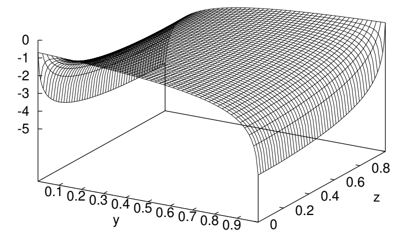

Figure 3: Function g(x,y,z) given by (17) with $x = 0.5$.

Therefore setting either $\theta_{2|1,1}$ or $\theta_{2|2,2}$ to 0.5 results in a smaller supremum of the log-likelihood than the above choices. Consequently, the four suprema are separated by areas where the log-likelihood is smaller, i.e., the suprema are local and not connected.

To conclude the proof we still need to address two issues: (a) the four local suprema give the same conditional log-likelihood $-2\ln 2$, and (b), they are suprema, not maxima (not achieved by any $\Theta^{\mathcal{B}} \in \mathbf{\Theta}^{\mathcal{B}}$). To deal with (a), consider data D' consisting of n_1 repetitions of $(1,1,1)$, n_2 repetitions of $(1,1,2)$, n_3 repetitions of $(2,2,1)$ and n_4 repetitions of $(2,2,2)$. By doing a slightly more involved analysis, one can show that, for some choices of n_1, n_2, n_3, n_4, the supervised log-likelihood still has four suprema, but they have different likelihood values. To deal with (b), let D'' be equal to D' but with four extra data vectors $(1,2,1),(2,1,1),(1,2,2),(2,1,2)$. If n_1, n_2, n_3 and n_4 are chosen large enough, the supervised likelihood for D'' has four maxima (rather than suprema), not all of which achieve the same supervised likelihood. We omit further details. ∎

Lemma 1. *With $0 < x < 1$ fixed and y and z both varying between 0 and 1, the supremum of $g(x,y,z)$ defined by* (17) *is given by*

$$\sup_{0 \le y,z \le 1} g(x,y,z) = \ln(x). \tag{20}$$

The function approaches its supremum when $z \in \{0,1\}$, and $y \to z$. That is, $\lim_{y \downarrow 0} g(x,y,0) = \lim_{y \uparrow 1} g(x,y,1) = \ln x$.

Proof. Differentiating twice wrt. z gives

$$\begin{aligned}
\frac{\partial^2}{\partial^2 z} g(x,y,z) &= \frac{(1-x)^2}{(xy + (1-x)z)^2} \\
&\quad + \frac{(1-x)^2}{(x(1-y) + (1-x)(1-z))^2},
\end{aligned} \tag{21}$$

which is always positive and the function achieves its maximum values at $z \in \{0,1\}$. At these two points derivating wrt. y yields

$$\begin{aligned}
\frac{\partial}{\partial y} g(x,y,0) &= \frac{x-1}{(1-y)(1-xy)}, \\
\frac{\partial}{\partial y} g(x,y,1) &= \frac{1-x}{y(xy+1-x)}.
\end{aligned} \tag{22}$$

Since in the first case the derivative is always negative, and in the second case the derivative is always positive, $g(x,y,0)$ increases monotonically as $y \to 0$, and $g(x,y,1)$ increases monotonically as $y \to 1$. In both cases the limiting value is $\ln(x)$. ∎