Bayesian Network Sensitivity to Arc-Removal

Silja Renooij
Department of Information and Computing Sciences, Utrecht University
P.O. Box 80.089, 3508 TB Utrecht, The Netherlands
silja@cs.uu.nl

Abstract

Arc-removal is usually employed as part of an approximate inference scheme for Bayesian networks, whenever exact inference is intractable. We consider the removal of arcs in a different setting, as a means of simplifying a network under construction. We show how sensitivity functions, capturing the effects of parameter variation on an output of interest, can be employed to describe detailed effects of removing an arc. In addition, we provide new insights related to the choice of parameter settings upon arc removal, and the effect of this choice on the quality of the simplified model as an approximation of the original one.

1 Introduction

Arc removal is a model simplification technique most often employed as part of an approximate inference scheme for Bayesian networks. Whenever exact inference is intractable, a set of "weak" links is selected and removed to arrive at an approximate network in which exact inference is feasible (Kjærulff, 1994; Engelen, 1997; Choi, Chan & Darwiche, 2005).

In this paper, we consider the removal of arcs in a different setting, where we are interested in simplifying a model that is being constructed with the help of domain experts. A sparser model has the computational benefits of having fewer arcs and, hence, a smaller number of probability parameters. Furthermore, a sparser model may be easier to understand for the domain experts. Our focus is now on gaining detailed insight into the possible impact of removing a single, pre-selected arc on the behaviour of the network, both with and without evidence.

We will demonstrate that by interpreting arc removal as a constrained form of varying multiple parameters in the original network, we can study the effects of such a removal by means of so-called *sensitivity functions*. More specifically, a sensitivity-to-arc-removal function will describe the effects of *any* choice of setting the new parameters after arc removal, on some output probability of interest. For establishing these functions, we assume that inference in the original network is possible.

The quality of the new network can be assessed by evaluating its behaviour. Alternatively, when the quality of the new network as an *approximation* of the original one is of concern, the sensitivity-to-arc-removal functions can be plugged into some quality measure. In the context of arc removal, the quality of the approximation is usually measured in terms of the KL-divergence between the prior joint distributions of the original network and the approximate network (Kjærulff, 1994; Engelen, 1997). The KL-divergence has the convenient property that the change in prior joint distribution occasioned by the removal of an arc $A \rightarrow B$ can be computed locally from the probabilities for variable B and its parents (Kjærulff, 1994). This property does not necessarily hold, however, if we consider the KL-divergence between marginal distributions, or between posterior distributions. We will show that in this setting, such KL-divergences can be computed with our sensitivity functions.

The paper is structured as follows. Section 2 briefly reviews Bayesian networks and sensitivity functions. In Section 3 we derive the sensitivity-to-arc-removal functions. Section 4 demonstrates the use of these functions for computing KL-divergence; in doing so, some novel insights into the effect on KL-divergence of different choices of new parameters are given. The paper ends with conclusions and directions for future research in Section 5.

2 Preliminaries

A Bayesian network compactly represents a joint probability distribution $\Pr$ over a set of stochastic variables $\mathbf{W}$ (Jensen & Nielsen, 2007). It combines an acyclic directed graph G, that captures the variables and their dependencies as nodes and arcs respectively, with conditional probability distributions $\Theta_{W_i|\pi(W_i)}$ for each variable W_i and its parents $\pi(W_i)$ in the graph, such that $\Pr(\mathbf{W}) = \prod_i \Theta_{W_i|\pi(W_i)}$.

We will refer to $\Theta_{W_i|\pi(W_i)}$ as the conditional probability table (CPT) of W_i; entries θ of Θ are called parameter probabilities, or parameters for short. In the remainder of this paper we will assume all variables to be binary-valued. Variables are denoted by capital letters and their values or instantiations by lower case; bold face is used for sets.

Probabilities computed from a Bayesian network are affected by the inaccuracies in the network's parameters. To investigate the extent of these effects, a sensitivity analysis can be performed in which $n \geq 1$ network parameters are varied simultaneously and the effect on an output probability of interest is studied. The effects of so-called n-*way* parameter variation are described by a *sensitivity function*. This is a multilinear function in the varied parameters in case of a prior probability of interest, and a rational function in the posterior case (Coupé & Van der Gaag, 2002). For example, the 2-*way* sensitivity function $f_{\Pr(a|\mathbf{e})}(x, y)$ describing the posterior probability $\Pr(a \mid \mathbf{e})$ as a function of two parameters x and y is given by

$$\frac{f_{\Pr(a\,\mathbf{e})}(x, y)}{f_{\Pr(\mathbf{e})}(x, y)} = \frac{c^{11}xy + c^{01}x + c^{10}y + c^{00}}{d^{11}xy + d^{01}x + d^{10}y + d^{00}}$$

where the constants c^{ij}, d^{ij}, $i, j \in \{0, 1\}$, are built from the non-varied parameters[1] in the network under study; feasible algorithms are available for their computation (Kjærulff & Van der Gaag, 2000; Coupé & Van der Gaag, 2002). Parameters from the same CPT $\Theta_{W_i|\pi(W_i)}$, but for different conditioning contexts, are independent; this results in zero interaction terms (Chan & Darwiche, 2004). In the above example this entails that $c^{11} = d^{11} = 0$.

[1] When a parameter θ varies as x, its complement $\overline{\theta} = 1 - \theta$ from the same distribution varies as $1 - x$. If θ concerns an k-valued variable, $k > 2$, then the $k - 1$ complementing parameters are co-varied proportionally.

3 Sensitivity Functions for Arc Removal

Sensitivity analysis typically refers to the study of effects of changes in network parameters on some outcome of interest. We can, however, also exploit it to study the effects of structural changes to the network's digraph, such as the removal of arcs.

Arc removal is most often employed as part of an approximate inference scheme, where arcs are removed until an approximate network is obtained in which exact inference is feasible (Kjærulff, 1994; Engelen, 1997; Choi, Chan & Darwiche, 2005). In this paper, we consider the removal of arcs in a different setting. We assume that we are constructing a Bayesian network with the help of domain experts, who are known to have the tendency of adding too many arcs into the model (Van der Gaag & Helsper, 2002). Our focus now is on studying the effects of removing a single arc, which we suspect may be superfluous, for the purpose of arriving at a simpler model that still suffices for the domain of application. We assume that inference in the original network is possible and, since we are still in a construction phase, that we have ample time to spend on it.

In this section, we propose the first approach for exactly studying the possible effects of arc removal on a probability of interest; the approach exploits the sensitivity function describing this probability in relation to the new parameters.

3.1 Implementing Arc Removal

Throughout this paper we consider the removal of an arc $A \rightarrow B$ from a Bayesian network $\mathcal{B}$, where $\pi(B) = \{A\} \cup \mathbf{Z}$. Removing an arc can be implemented in various ways (Choi, Chan & Darwiche, 2005). The approach we adopt in this paper is to simulate the removal by changes in the CPT $\Theta_{B|A\mathbf{Z}}$. More specifically, for each combination of values b and $\mathbf{z}$ the parameters $\theta_{b|a\mathbf{z}}$ are set to be equal for all values a; we will refer to these new parameters as $\theta'_{b|.\mathbf{z}}$, since the value of A is irrelevant.

The more or less standard approach to setting the new parameter values is by marginalising out variable A, or by approximating this process if it is infeasible to perform the exact computations (Engelen, 1997; Choi, Chan & Darwiche, 2005). More recent approaches focus on the addition of auxiliary nodes and parameters that compensate for the

lost dependency, together with iterative or variational methods that optimise these parameters as part of the arc-removal procedure (Choi & Darwiche, 2010; Choi & Darwiche, 2006). Rather than choosing a new parameter setting in advance, however, we can study the effects of all possible settings.

3.2 Sensitivity to Arc Removal

We study the effects of arc removal using a sensitivity analysis in which we vary all parameters $\theta_{b|a\mathbf{z}}$ until, for each b and context $\mathbf{z}$, the parameters are equivalent for all a. Let m be the number of different instantiations $\mathbf{z}$ for $\mathbf{Z}$, then arc removal requires the simultaneous variation of $2m$ parameters[2]. Generally, determining a $2m$-way sensitivity function is computationally demanding. The following proposition shows, however, that in the context of arc removal an m-way function suffices. Moreover, this function can be obtained from $2m$ 1-way sensitivity functions, which *can* be established efficiently (Kjærulff & Van der Gaag, 2000).

Proposition 1. *Let i index the values of variable A and j the instantiations of $\mathbf{Z}$. Let $x_{ij} = \theta_{b_1|a_i\mathbf{z}_j}$ and $1 - x_{ij} = \theta_{b_2|a_i\mathbf{z}_j}$ denote the $2 \cdot 2m$ parameters in $\Theta_{B|A\mathbf{Z}}$, and let $\Theta'_{B|.\mathbf{Z}}$ be the result of setting $x_{1j} = x_{2j}$ for all j. Let x'_j and $1-x'_j$ denote the parameters in Θ'. Then, the m-way sensitivity function which captures the effects of any possible choice for the parameters in $\Theta'_{B|.\mathbf{Z}}$ on an output probability of interest $\Pr(\mathbf{v})$ equals:*

$$f_{\Pr(\mathbf{v})}(x'_1, \ldots, x'_m) = \Big(\sum_{j=1}^{m}\Big(\sum_{i=1}^{2} c^1_{ij}\Big)\cdot x'_j\Big) + \\ + \Big(\sum_{i=1}^{2}\sum_{j=1}^{m} c^0_{ij}\Big) - (2\cdot m - 1)\cdot \Pr(\mathbf{v})$$

where c^0_{ij} and c^1_{ij} equal the constants from the 1-way sensitivity function describing $\Pr(\mathbf{v})$ as a function of parameter x_{ij}, $f_{\Pr(\mathbf{v})}(x_{ij}) = c^1_{ij}\cdot x_{ij} + c^0_{ij}$.

Proof: First we will detail the probabilistic semantics of the constants of a sensitivity function for $\Pr(\mathbf{v})$. Each of the $2m$ terms in the summation

$$\Pr(\mathbf{v}) = \sum_{i=1}^{2}\sum_{j=1}^{m} \Pr(\mathbf{v}\, a_i\mathbf{z}_j)$$

depends only on parameters $\theta_{B|a_i\mathbf{z}_j}$ with a corresponding conditioning context, and is constant with respect to other parameters in the CPT of B. More specifically, each $\Pr(\mathbf{v}\, a_i\mathbf{z}_j)$ relates to parameter $\theta_{b_1|a_i\mathbf{z}_j}$ as follows:

$$\begin{aligned}\Pr(\mathbf{v}\, a_i\mathbf{z}_j) &= \\ &= \sum_{k=1}^{2} \Pr(\mathbf{v} \mid b_k a_i\mathbf{z}_j)\cdot\Pr(b_k \mid a_i\mathbf{z}_j)\cdot\Pr(a_i\mathbf{z}_j) \\ &= \Pr(\mathbf{v} \mid b_1 a_i\mathbf{z}_j)\cdot\theta_{b_1|a_i\mathbf{z}_j}\cdot\Pr(a_i\mathbf{z}_j) \\ &\quad + \Pr(\mathbf{v} \mid b_2 a_i\mathbf{z}_j)\cdot(1-\theta_{b_1|a_i\mathbf{z}_j})\cdot\Pr(a_i\mathbf{z}_j)\end{aligned}$$

$\Pr(\mathbf{v})$ in terms of a single $\theta_{b_1|a_i\mathbf{z}_j}$ thus equals

$$\Pr(\mathbf{v}) = \big(c^1_{ij}\cdot\theta_{b_1|a_i\mathbf{z}_j} + r_{ij}\big) + \big(\Pr(\mathbf{v}) - \Pr(\mathbf{v}\, a_i\mathbf{z}_j)\big)$$

where

$$c^1_{ij} = \Big(\Pr(\mathbf{v} \mid b_1 a_i\mathbf{z}_j) - \Pr(\mathbf{v} \mid b_2 a_i\mathbf{z}_j)\Big)\cdot\Pr(a_i\mathbf{z}_j)$$

and $r_{ij} = \Pr(\mathbf{v} \mid b_2 a_i\mathbf{z}_j)\cdot\Pr(a_i\mathbf{z}_j)$. For $\Pr(\mathbf{v})$ in relation to $x_{ij} = \theta_{b_1|a_i\mathbf{z}_j}$ we therefore have a 1-way sensitivity function of the form $f_{\Pr(\mathbf{v})}(x_{ij}) = c^1_{ij}\cdot x_{ij} + c^0_{ij}$ with

$$c^0_{ij} = r_{ij} + \Pr(\mathbf{v}) - \Pr(\mathbf{v} a_i\mathbf{z}_j)$$

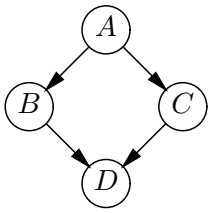
Consequently, upon varying all $2m$ parameters $x_{1j} = \theta_{b_1|a_1\mathbf{z}_j}$ and $x_{2j} = \theta_{b_1|a_2\mathbf{z}_j}$, $j = 1, \ldots, m$, we find from

$$\begin{aligned}\Pr(\mathbf{v}) &= \sum_{i=1}^{2}\sum_{j=1}^{m}\Big(c^1_{ij}\cdot\theta_{b_1|a_i\mathbf{z}_j} + r_{ij}\Big) \\ &= \sum_{i=1}^{2}\sum_{j=1}^{m}\Big(c^1_{ij}\cdot\theta_{b_1|a_i\mathbf{z}_j} + c^0_{ij}\Big) \\ &\quad - \sum_{i=1}^{2}\sum_{j=1}^{m}\Big(\Pr(\mathbf{v}) - \Pr(\mathbf{v}\, a_i\mathbf{z}_j)\Big)\end{aligned}$$

the function, $f_{\Pr(\mathbf{v})}(x_{11}, x_{21}, \ldots, x_{1m}, x_{2m}) =$

$$\begin{aligned}&= \Big(\sum_{i=1}^{2}\sum_{j=1}^{m} f_{\Pr(\mathbf{v})}(x_{ij})\Big) - (2\cdot m - 1)\cdot\Pr(\mathbf{v}) \\ &= \sum_{i=1}^{2}\sum_{j=1}^{m} c^1_{ij}\cdot x_{ij} + c^0\end{aligned}$$

[2] Another $2m$ parameters, for the other value of B, are covaried.



$\theta_{a_1} = 0.8$

$\theta_{b_1|a_1} = 0.8$
$\theta_{b_1|a_2} = 0.4$

$\theta_{c_1|a_1} = 0.5$
$\theta_{c_1|a_2} = 1.0$

$\theta_{d_1|b_1c_1} = 0.1$
$\theta_{d_1|b_1c_2} = 0.3$
$\theta_{d_1|b_2c_1} = 0.9$
$\theta_{d_1|b_2c_2} = 0.8$

Figure 1: The example Bayesian network, taken from Choi & Darwiche (2010).

with $c^0 = \Big(\sum_{i=1}^{2}\sum_{j=1}^{m} c^0_{ij}\Big) - (2{\cdot}m - 1){\cdot}\Pr(\mathbf{v})$

The above $2m$-way sensitivity function describes all possible effects of varying the $2m$ parameters on $\Pr(\mathbf{v})$. To study the effects of arc removal, however, variation of these parameters is constrained in the sense that $x_{1j} = x_{2j}$ should hold for each $j = 1, \ldots, m$. That is, rather than a $2m$ dimensional parameter space, we are actually dealing with an m dimensional space and an m-way sensitivity function. Using x'_j to denote the parameters $\theta'_{b_1|.\mathbf{z}_j}$ resulting from setting $x_{1j} = x_{2j}$, we conclude

$$f_{\Pr(\mathbf{v})}(x'_1, \ldots, x'_m) = \Big(\sum_{j=1}^{m}\Big(\sum_{i=1}^{2} c^1_{ij}\Big){\cdot}x'_j\Big) + c^0$$

□

Note that in the above proposition we have not assumed $\Pr(\mathbf{v})$ to be a marginal over a single variable. The proposition is therefore more generally applicable, but most algorithms for computing the constants of sensitivity functions assume that we are interested in a single-variable marginal (prior or posterior). Generalisation of the proposition to a posterior probability of interest, such as $\Pr(\mathbf{v} \mid \mathbf{e}) = \frac{\Pr(\mathbf{ve})}{\Pr(\mathbf{e})}$, is also straightforward, as we demonstrate in the following example.

Example 1. Consider the Bayesian network in Figure 1, which will serve as a running example throughout the paper. We assume that variable A can take on two values, represented by a_1 and a_2, respectively. A similar assumption holds for the other variables in the network. We are interested in studying the effects of removing arc $A \rightarrow B$, and therefore consider the parameters $x_1 = \theta_{b_1|a_1}$ and $x_2 = \theta_{b_1|a_2}$, and their complements.

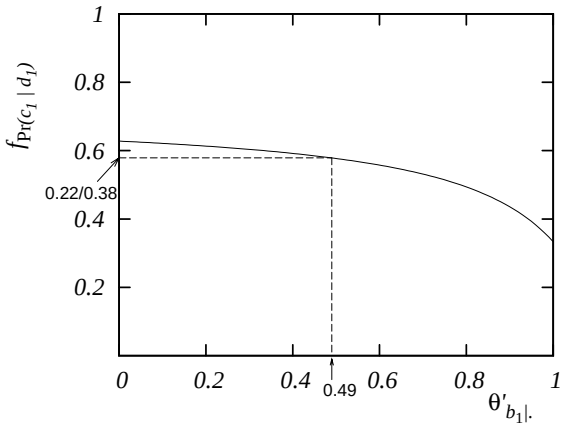


Figure 2: $\Pr(c_1 \mid d_1)$ as a function of the new parameters $\theta'_{b_1|.}$ and $\theta'_{b_2|.} = 1 - \theta'_{b_1|.}$.

Suppose $\Pr(c_1 \mid d_1)$ is our probability of interest, with original value $\Pr(c_1 d_1)/\Pr(d_1) = 0.22/0.38$. The relevant 1-way sensitivity functions are:

$$f_{\Pr(c_1|d_1)}(x_1) = \frac{-0.32{\cdot}x_1 + 0.48}{-0.52{\cdot}x_1 + 0.80}$$

$$f_{\Pr(c_1|d_1)}(x_2) = \frac{-0.16{\cdot}x_2 + 0.28}{-0.16{\cdot}x_2 + 0.44}$$

Following Proposition 1, simultaneous variation of x_1 and x_2 results in: $f_{\Pr(c_1|d_1)}(x_1, x_2) =$

$$= \frac{-0.32{\cdot}x_1 - 0.16{\cdot}x_2 + 0.48 + 0.28 - 0.22}{-0.52{\cdot}x_1 - 0.16{\cdot}x_2 + 0.80 + 0.44 - 0.38}$$

$$\Longrightarrow \frac{-0.48{\cdot}x' + 0.54}{-0.68{\cdot}x' + 0.86} = f_{\Pr(c_1|d_1)}(x')$$

where x' results from setting $x_1 = x_2$. This function, shown in Figure 2, now describes the possible effects of removing arc $A \rightarrow B$ on the probability $\Pr(c_1 \mid d_1)$. From the function we can, for example, compute which new parameter setting will result in the original value of the probability of interest:

$$f_{\Pr(c_1|d_1)}(x') = \frac{0.22}{0.38} \iff x' = 0.49 \qquad \square$$

Although we assumed all variables to be binary, Proposition 1 trivially extends to non-binary variables A and $\mathbf{X}$. If variable B can take on $n > 2$ values, however, the sensitivity function describing the effects of arc removal can no longer be obtained from 1-way sensitivity functions. In that case, varying the parameters $\theta_{b_1|a_i\mathbf{z}}$ until they become equal for all a_i, no longer ensures that all proportionally

co-varying parameters $\theta_{b_2|a_i\mathbf{z}} \ldots \theta_{b_n|a_i\mathbf{z}}$, $n > 2$, become equal for all a_i. To enforce such equalities, $(n-1)$-way analyses are necessary.

By studying the sensitivity functions that describe the effects of arc removal on an output of interest for *various outputs* and for *various combinations of observations*, we can determine whether there exists a parameter setting that results in acceptable behaviour of the simplified model. From the sensitivity functions we can, for example, immediately determine if a specific case entered into the network would result in the same most likely value of our output variable in both the original and the simplified network. Another way to compare the models, is by comparing the distributions they define.

4 Arc Removal and KL-divergence

Empirical evidence shows that arc removal can lead to quite a speedup in inference, at the cost of only little deterioration in quality (Choi, Chan & Darwiche, 2005; Santana & Provan, 2008; Choi & Darwiche, 2010). In this context, quality is measured in terms of the Kullback-Leibler (KL) divergence between the original distribution $\Pr$ and the distribution $\Pr'$ for the approximate network (Cover & Thomas, 1991); it is defined by

$$\mathrm{KL}(\Pr(\mathbf{V}), \Pr'(\mathbf{V})) \stackrel{\mathbf{def}}{=} \sum_{\mathbf{v}} \Pr(\mathbf{v}) \cdot \log \frac{\Pr(\mathbf{v})}{\Pr'(\mathbf{v})}$$

The KL-divergence has the convenient property that the change in *prior joint* distribution occasioned by the removal of an arc $A \rightarrow B$ can be computed locally from the probabilities for variable B and its parents (Kjærulff, 1994). This property does not necessarily hold, however, if we consider the KL-divergence between marginal distributions, or between posterior distributions, which are typically of interest for practical applications. Recently, it was argued that arc-removal methods should take available evidence into account, in order to tailor the approximation to the evidence at hand (Choi, Chan & Darwiche, 2005; Choi & Darwiche, 2010). In this section we will consider KL-divergence as a function of the new parameter settings, and demonstrate that we can plug in our sensitivity-to-arc-removal functions in order to compute this divergence, both between joint and marginal distributions, with and without evidence.

4.1 Joint Prior and Joint Posterior Divergence

If, upon arc removal, we wish to choose the new parameter settings such that they minimise the KL-divergence between the original and the simplified network, then under some conditions this choice is evident. The clear-cut cases concern joint (prior or posterior) distributions and are given by the proposition below.

In the remainder of this section we let network $\mathcal{B}'$ be the result of removing arc $A \rightarrow B$ from the original network $\mathcal{B}$. The distribution defined by $\mathcal{B}'$ is denoted $\Pr'$.

Proposition 2. *Consider the two joint prior distributions* $\Pr(\mathbf{V})$ *and* $\Pr'(\mathbf{V})$*, and two joint posterior distributions* $\Pr(\mathbf{V} \mid \mathbf{e})$ *and* $\Pr'(\mathbf{V} \mid \mathbf{e})$ *conditioned on evidence* $\mathbf{e}$*. Then*

- $\mathrm{KL}(\Pr(\mathbf{V}), \Pr'(\mathbf{V}))$ *is minimised by setting, for all* b *and* $\mathbf{z}$ *combinations,* $\theta'_{b|.\mathbf{z}} = \Pr(b \mid \mathbf{z})$*;*
- $\mathrm{KL}(\Pr(\mathbf{V} \mid \mathbf{e}), \Pr'(\mathbf{V} \mid \mathbf{e}))$ *is minimised by setting, for all* b *and* $\mathbf{z}$ *combinations,* $\theta'_{b|.\mathbf{z}} = \Pr(b \mid \mathbf{z}\mathbf{e})$*, if* $\Pr(\mathbf{e}) = \Pr'(\mathbf{e})$.

Proof: The factorisation of the joint distribution is exploited to reduce the KL-divergence to terms involving the CPT of variable B (see Kjærulff (1994) or Engelen (1997) for the prior situation and Choi, Chan & Darwiche (2005) for the posterior case):

$$\begin{aligned}
&\mathrm{KL}(\Pr(\mathbf{V} \mid \mathbf{e}), \Pr'(\mathbf{V} \mid \mathbf{e})) = \\
&\quad = \sum_{\mathbf{v}} \Pr(\mathbf{v} \mid \mathbf{e}) \cdot \log \frac{\Pr(\mathbf{v} \mid \mathbf{e})}{\Pr'(\mathbf{v} \mid \mathbf{e})} \\
&\quad = \log \frac{\Pr'(\mathbf{e})}{\Pr(\mathbf{e})} + \sum_{ab\mathbf{z}} \Pr(ab\mathbf{z} \mid \mathbf{e}) \cdot \log \frac{\theta_{b|a\mathbf{z}}}{\theta'_{b|.\mathbf{z}}}
\end{aligned}$$

The term $\sum_{ab\mathbf{z}} \Pr(ab\mathbf{z} \mid \mathbf{e}) \cdot \log \theta_{b|a\mathbf{z}}$ is determined by the original network only. The remaining terms are a function of the new parameters and equals:

$$\begin{aligned}
&\log \frac{\Pr'(\mathbf{e})}{\Pr(\mathbf{e})} - \sum_{ab\mathbf{z}} \Pr(ab\mathbf{z} \mid \mathbf{e}) \cdot \log \theta'_{b|.\mathbf{z}} = \\
&= \log \frac{\Pr'(\mathbf{e})}{\Pr(\mathbf{e})} + \sum_{\mathbf{z}} \Pr(\mathbf{z} \mid \mathbf{e}) \cdot \\
&\qquad \cdot \Big(- \sum_{b} \Pr(b \mid \mathbf{z}\mathbf{e}) \cdot \log \theta'_{b|.\mathbf{z}} \Big)
\end{aligned}$$

The bracketed summation equals the cross-entropy between the two distributions over B and is known to be minimal if the distributions are the same, i.e. $\theta'_{b|.\mathbf{z}} = \Pr(b \mid \mathbf{ze})$.

In the prior situation, we get the same formula but without the $\log(\Pr'(\mathbf{e})/\Pr(\mathbf{e}))$ term, and with the $\mathbf{e}$'s removed from the conditioning contexts. In that case, minimising cross-entropy, i.e. setting $\theta'_{b|.\mathbf{z}} = \Pr(b \mid \mathbf{z})$, serves to minimise the KL-divergence. In the posterior case, however, minimising cross-entropy is only *guaranteed* to minimise the KL-divergence if $\Pr'(\mathbf{e}) = \Pr(\mathbf{e})$, i.e. if the probability of evidence is insensitive to changes in the parameters for B. □

The first property in the above proposition, although to the best of our knowledge never explicitly proven, must be well-known: the optimal parameter setting stated amounts exactly to marginalising out variable A, which is a standard approach to implementing arc removal.

For the two cases stated in Proposition 2, we have an expression defining the KL-divergence in relation to the new parameter settings. In case $\Pr(\mathbf{e}) \neq \Pr'(\mathbf{e})$, we can now plug in the sensitivity function $f_{\Pr'(\mathbf{e})}(\theta'_{b|.\mathbf{z}})$ and again get the KL-divergence as a function of the new parameters. For low-dimensional functions it is then easy to compute the parameter settings that minimise the divergence.

Example 2. Reconsider the example Bayesian network in Figure 1. We will use the terms *prior KL-divergence* and *posterior KL-divergence* to refer to the divergence between (joint) prior and posterior distributions, respectively. The prior KL-divergence as a function of $x' = \theta'_{b_1|.}$ equals

$$\begin{aligned}\mathrm{KL}(\Pr(ABCD), \Pr'(ABCD))(x') = \\ = \sum_{i=1}^{2}\sum_{j=1}^{2} \Pr(a_i b_j)\cdot\log\theta_{b_j|a_i} + \\ - \Pr(b_1)\cdot\log x' - \Pr(b_2)\cdot\log(1-x') \\ = -0.77 - 0.72\cdot\log x' - 0.28\cdot\log(1-x')\end{aligned}$$

and is shown in Figure 3 (dashed). We can indeed verify from the figure that the values of the new parameters that correspond with the marginal probabilities for B, i.e. $\theta'_{b_1|.} = 0.80\cdot0.8 + 0.4\cdot0.2 = 0.72$ and $\theta'_{b_2|.} = 0.28$, result in a minimal KL-divergence

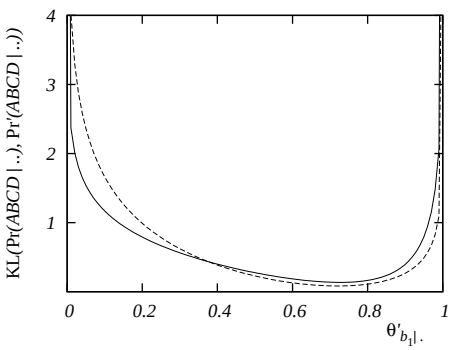


Figure 3: The prior KL-divergence (dashed), and the posterior divergence, given evidence d_1, as a function of the new parameters $\theta'_{b_1|.}$ and $\theta'_{b_2|.} = 1 - \theta'_{b_1|.}$ (zero and one excluded).

of 0.08. Figure 3 also shows the joint *posterior* KL-divergence in the context of evidence d_1 (solid); this can be written in terms of x' by using the sensitivity function $f_{\Pr'(d_1)}(x')$:

$$\begin{aligned}\mathrm{KL}(\Pr(ABC \mid d_1), \Pr'(ABC \mid d_1))(x') = \\ = \log\frac{-0.68\cdot x' + 0.86}{0.38} + 0.52 \\ - 0.28\cdot\log x' - 0.24\cdot\log(1-x')\end{aligned}$$

Taking the first derivative, we find that this function is minimised for $x' = 0.73$. This same value for $\Pr(b_1)$ follows from the auxiliary parameters established, for this *same* example network, by the iterative procedure in Choi & Darwiche (2010). Note that this optimal value is found for a much higher value of x' than would be found through marginalisation, i.e. $\theta'_{b|.\mathbf{z}} = \sum_a \Pr(b \mid a\mathbf{ze})\cdot\Pr(a \mid \mathbf{ze})$:

$$\theta'_{b_1|.} = 0.48\cdot0.69 + 0.07\cdot0.31 = 0.36$$

This is caused by the fact that $\Pr(d_1)$ is sensitive to the parameter changes. The parameter setting used in (Choi, Chan & Darwiche, 2005), $\theta'_{b|.\mathbf{z}} = \sum_a \theta_{b|a\mathbf{z}}\cdot\Pr(a \mid \mathbf{e})$, results in:

$$\theta'_{b_1|.} = 0.8\cdot0.69 + 0.4\cdot0.31 = 0.68$$

Remarkably, these settings are closer to the optimum, despite their use of the invalid independence assumption that B is independent of D given A. □

4.2 Marginal Prior and Posterior Divergence

Suppose we wish to choose the new parameter settings such that they minimise the KL-divergence between marginal rather than joint distributions. The

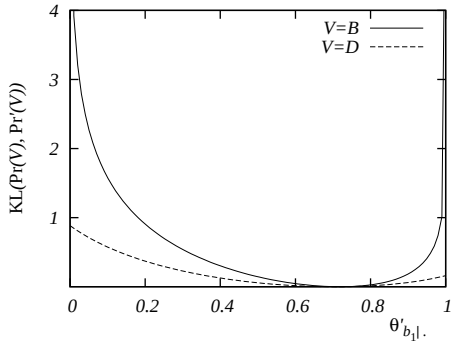


Figure 4: Prior marginal KL-divergences, per variable, as a function of the new parameters $\theta'_{b_1|.}$ and $\theta'_{b_2|.} = 1 - \theta'_{b_1|.}$. For variables A and C the divergence is zero.

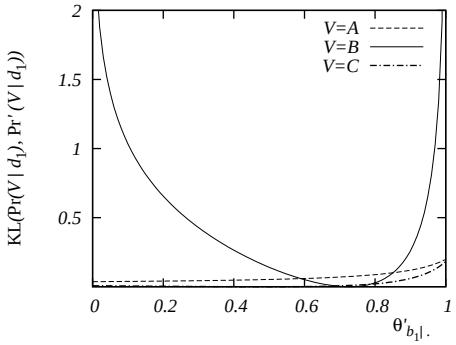


Figure 5: Posterior marginal KL-divergences per variable, given evidence d_1, as a function of the new parameters $\theta'_{b_1|.}$ and $\theta'_{b_2|.} = 1 - \theta'_{b_1|.}$.

KL-divergence between marginal distributions is hardly ever considered, since in that case we cannot exploit the factorisation of the joint distribution to reduce the divergence to local terms. Using the sensitivity functions for arc removal introduced in the previous section, however, we can define the KL-divergence as a function of the new parameters under consideration.

Corollary 1. *Let* $\Pr$, $\Pr'$ *and* $x'_1, \ldots x'_m$ *be as before, then*

$$\mathrm{KL}(\Pr(\mathbf{V} \mid \mathbf{e}), \Pr'(\mathbf{V} \mid \mathbf{e}))(x'_1, \ldots x'_m) =$$

$$= \sum_{\mathbf{v}} \Pr(\mathbf{v} \mid \mathbf{e}) \cdot \log \frac{\Pr(\mathbf{v} \mid \mathbf{e})}{f_{\Pr'(\mathbf{v}|\mathbf{e})}(x'_1, \ldots x'_m)}$$

Note that the above corollary applies to both joint and marginal distributions; in the prior situation, the above holds with all occurrences of $\mathbf{e}$ removed. Its formula in fact was used to create the graphs of Figures 3, 4 and 5.

The following example illustrates, for each variable V in our example network, the KL-divergence between the original marginal distribution $\Pr(V)$ and the new marginal distribution $\Pr'(V)$, for different choices of the new parameters for variable B. We will refer to these divergences as *marginal KL-divergences*.

Example 3. Reconsider the example Bayesian network in Figure 1, from which we remove arc $A \rightarrow B$. Figure 4 now shows the prior marginal KL-divergences for each variable and all possible choices for the new parameters θ'_{b_i} (zero and one excluded). Since the values in the CPT for variable B affect only the marginal distributions of B and D, the KL-divergences for A and C are zero. In addition, we see that the parameter settings that optimise the marginal KL-divergence for B, also optimise the divergence for its descendant D.

Figure 5 similarly shows the various marginal KL-divergences in the context of evidence d_1. We see that the parameter setting for which the KL-divergence is optimal, now varies per variable: this is due to the fact that the marginal KL-divergence for a certain variable is optimal, if the new parameters are chosen such that the new marginal probabilities equal the original ones.

As an illustration, suppose we are interested in variable C, in the context of evidence d_1. Recall that the sensitivity function for c_1 is given by

$$f_{\Pr'(c_1|d_1)}(x') = \frac{-0.48 \cdot x' + 0.54}{-0.68 \cdot x' + 0.86}$$

The original value for $\Pr(c_1 \mid d_1)$ equals $\frac{0.22}{0.38}$, which we can obtain in our new network by setting $x' = \theta'_{b_1|.} = 0.49$. Figure 5 indeed suggests that this choice is optimal in terms of KL-divergence. □

From the previous examples we have that the optimal choice for the new parameter settings, in terms of minimising the KL-divergence, differs between prior and posterior distributions, both for joint and marginal distributions. In the example network, however, setting the new parameter $\theta'_{b_1|.}$ to a value somewhere in the range $[0.6, 0.8]$ seems to result in a small KL-divergence in all situations considered.

5 Conclusion

In this paper we introduced sensitivity functions as a means of studying the exact impact of removing an arc from a Bayesian network on some output of interest. These functions provide insight into whether or not removing the arc can result in an acceptable simplified model, and they can support our choice for setting the new parameters upon removal. If the simplified network should be a good quality approximation of the original one, then the sensitivity functions can also be used to find new parameters that minimise the KL-divergence between various distributions of interest.

In addition, we provided some insights concerning arc removal and KL-divergence. More specifically, we showed that arc removal by means of marginalisation is in fact optimal in terms of minimising the KL-divergence between prior joint distributions. Secondly, we provided a condition under which marginalisation results in an optimal KL-divergence between posterior joint distributions.

We assumed that all variables are binary-valued. As mentioned, extension to non-binary variables is trivial, except for variable B. For non-binary B, proportional co-variation of its values no longer ensures that all parameters that should be equated for arc removal in fact are. As a result, multi-way functions with non-zero interaction terms are necessary. Further research is required to establish the exact implications of this increased complexity.

Although our interest in arc removal is not in approximating networks to make inference feasible, our results can be put to use in situations where the complexity of a network is such that exact inference is possible, but too time-consuming for practical purposes. In such a case, detailed insights concerning the effects of arc removal can be obtained prior to deploying the network, and then exploited to construct efficient approximations for certain sets of observable variables, or for varying output variables of interest.

Acknowledgement

I would like to thank Linda van der Gaag for our inspiring discussions and the anonymous reviewers for useful comments.

References

H. Chan, A. Darwiche (2004). Sensitivity analysis in Bayesian networks: from single to multiple parameters. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, AUAI Press, pp. 67 – 75.

A. Choi, H. Chan, A. Darwiche (2005). On Bayesian network approximation by edge deletion. *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, AUAI Press, pp. 128 – 135.

A. Choi, A. Darwiche (2006). A variational approach for approximating Bayesian networks by edge deletion. *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, AUAI Press, pp. 80 – 89.

A. Choi, A. Darwiche (2010). An edge deletion semantics for belief propagation. *Submitted for publication*. Available as `http://reasoning.cs.ucla.edu/fetch.php?id=98&type=pdf`

V.M.H. Coupé, L.C. van der Gaag (2002). Properties of sensitivity analysis of Bayesian belief networks. *Annals of Mathematics and Artificial Intelligence*, 36, pp. 323 – 356.

Th.M. Cover, J.A. Thomas (1991). *Elements of Information Theory*, Wiley-Interscience.

R.A. van Engelen (1997). Approximating Bayesian belief networks by arc removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, pp. 916 – 920.

F.V. Jensen, T.D. Nielsen (2007). *Bayesian Networks and Decision Graphs*, Springer-Verlag.

U. Kjærulff (1994). Reduction of computational complexity in Bayesian networks through removal of weak dependencies. *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, pp. 374 – 382.

U. Kjærulff, L.C. van der Gaag (2000). Making sensitivity analysis computationally efficient. *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, pp. 317 – 325.

A. Santana, G. Provan (2008). An analysis of Bayesian network model-approximation techniques. *Proceedings of the 18th European Conference on Artificial Intelligence*, IOS Press, pp. 851 – 852.

L.C. van der Gaag, E.M. Helsper (2002). Experiences with modelling issues in building probabilistic networks. *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, LNCS 2473, Springer-Verlag, pp. 21 – 26.