Bayesian Networks: a Combined Tuning Heuristic

Janneke H. Bolt J.H.BOLT@UU.NL
Department of Information and Computing Sciences
Utrecht University P.O. Box 80.089, 3508 TB Utrecht

Abstract

One of the issues in tuning an output probability of a Bayesian network by changing multiple parameters is the relative amount of the individual parameter changes. In an existing heuristic parameters are tied such that their changes induce locally a maximal change of the tuned probability. This heuristic, however, may reduce the attainable values of the tuned probability considerably. In another existing heuristic parameters are tied such that they simultaneously change in the entire interval $\langle 0, 1 \rangle$. The tuning range of this heuristic will in general be larger then the tuning range of the locally optimal heuristic. Disadvantage, however, is that knowledge of the local optimal change is not exploited. In this paper a heuristic is proposed that is locally optimal, yet covers the larger tuning range of the second heuristic. Preliminary experiments show that this heuristic is a promising alternative.

Keywords: Bayesian networks; network tuning.

1. Introduction

Parameter tuning is one of the tools used for the construction of a Bayesian network that faithfully represents a problem domain of interest. An expected output of the network under construction then is the compared with the actual output and if the actual output deviates one or more of the network's parameters are adapted in order to enforce the correct output. In network tuning, changing multiple parameters may be preferred over a change of just a single parameter probability. In some cases simply because a desired effect is not attainable by a single parameter change; also a more levelled change of a network may be achieved when multiple parameters are used for tuning. When a network is tuned by changing multiple parameters, choices have to be made with respect to their relative amount of change; there may be several solutions and it may be preferred to use a combination of changes that disturbs the original distribution as little as possible.

In Bolt and Renooij (2014) a 'locally optimal' tuning heuristic was proposed. With this heuristic, provided that the required change of the output probability is small, approximately the solution involving the smallest parameter changes is found. A disadvantage of this method, however, is that the range of attainable values for the tuned probability can be reduced considerably. In the 'balanced' heuristic Chan and Darwiche (2004); Bolt and van der Gaag (2015), tuning parameters simultaneously change in the entire interval $\langle 0, 1 \rangle$ and this heuristic therefore covers in general a larger tuning range that the locally optimal heuristic. A disadvantage of the balanced heuristic, however, is that information about the locally optimal change of the parameters is not exploited. In Bolt and Renooij (2014) and Bolt and van der Gaag (2015) moreover sliced and balanced sensitivity functions were introduced with which the change of an output probability given the locally optimal and the balanced heuristic can be described.

In this paper a tuning heuristic is proposed that combines the advantages of locally optimal and the balanced heuristic. In this combined heuristic the parameter changes are locally optimal, yet, the heuristic covers the same tuning range as the balanced heuristic. In preliminary experiments the heuristics are evaluated using two different distance measures for probability distributions. The experiments show that the combined heuristic is a promising alternative for network tuning involving multiple parameters with respect to minimising the distance between the original and the tuned network.

2. Preliminaries

This paper considers Bayesian networks $\mathcal{B}$ representing joint probability distributions over sets of discrete random variables $\mathbf{V} = \{V_1, \ldots, V_n\}$, $n \geq 1$. A value of V is indicated by v and a joint value combination of $\mathbf{V}$ is indicated by $\mathbf{v}$. The capitals V and $\mathbf{V}$ are also used to indicate the set of values of V and the set of possible value combinations of $\mathbf{V}$. Given a binary variable, v and $\bar{v}$ are used to indicate its values. The letters x, y and z are used to indicate individual parameters of a network and $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$ are used to indicate sets of parameters.

When a parameter x of $\mathcal{B}$ is said to have a guaranteed positive effect on an outcome probability $\Pr(\mathbf{w} \mid \mathbf{u})$ of $\mathcal{B}$, this entails that increasing (decreasing) the value of x will increase (decrease) $\Pr(\mathbf{w} \mid \mathbf{u})$, whatever the values of the other, non-varied, parameters of $\mathcal{B}$. The meaning of a guaranteed negative effect is analogous. Parameters with a guaranteed positive or negative effect will said to have a guaranteed qualitative effect. The superscript ^{o} will be used to indicate parameter values and probabilities of a network in its original state.

2.1 N-way, Sliced and Balanced Sensitivity Functions

This section discusses n-way, sliced and balanced sensitivity functions. Sliced and balanced sensitivity functions can be considered as 'constrained' sensitivity functions which can be derived from an underlying n-way function. The locally optimal and the balanced tuning heuristic, which are discussed in Section 3, can be described by sliced and balanced sensitivity functions, respectively.

An n-way sensitivity function $\Pr(\mathbf{w} \mid \mathbf{u})(\mathbf{z})$ expresses an output $\Pr(\mathbf{w} \mid \mathbf{u})$ of a Bayesian network $\mathcal{B}$ in terms of n of its parameters, $\mathbf{z} = \{z_1, \ldots, z_n\}$ Coupé and van der Gaag (2002). More specifically, a higher-order sensitivity function in $\mathbf{z}$ takes the form of a fractional-multilinear function:

$$\Pr(\mathbf{w} \mid \mathbf{u})(\mathbf{z}) = \frac{\sum_{\mathbf{z}_k \in \mathcal{P}(\mathbf{z})} \left(c_k \cdot \prod_{z_i \in \mathbf{z}_k} z_i \right)}{\sum_{\mathbf{z}_k \in \mathcal{P}(\mathbf{z})} \left(d_k \cdot \prod_{z_i \in \mathbf{z}_k} z_i \right)}$$

with $\mathbf{U} \cap \mathbf{W} = \varnothing$, where $\mathcal{P}(\mathbf{z})$ is the powerset of the set parameters $\mathbf{z}$, and where the constants c_k, d_k are determined by (a subset of) the non-varied parameters of the network at hand. A two-way sensitivity function in the parameters z_1 and z_2 for example, has the following general form:

$$\Pr(\mathbf{w} \mid \mathbf{u})(z_1, z_2) = \frac{c_0 + c_1 \cdot z_1 + c_2 \cdot z_2 + c_3 \cdot z_1 \cdot z_2}{d_0 + d_1 \cdot z_1 + d_2 \cdot z_2 + d_3 \cdot z_1 \cdot z_2}$$

As is usual, it is supposed that at most one parameter per local distribution is actively varied and that the other parameters of the distribution, are adjusted according a proportional co-variation scheme. Moreover, in this paper, it is assumed that parameters with a value of 0 or 1 are not adapted, and that adapted parameters are not changed to 0 or 1.

If the sign of the partial derivative of $\Pr(\mathbf{w} \mid \mathbf{u})(\mathbf{z})$ with respect to $z_i \in \mathbf{z}$ is positive at $\mathbf{z}^o$, the increase of the parameter z_i has at $\mathbf{z}^o$ a positive effect on the output $\Pr(\mathbf{w} \mid \mathbf{u})$ and if it is negative, the increase of z_i has a negative effect. In the remainder $z_i \updownarrow z_j$ and $z_i \not\updownarrow z_j$ will be used to indicate that the partial derivatives of $\Pr(\mathbf{w}\,|\,\mathbf{u})(\mathbf{z})$ with respect to z_i and z_j have an equal, respectively, an opposite sign at $\mathbf{z}^o$. Note that the partial derivative of $\Pr(\mathbf{w}\,|\,\mathbf{u})(\mathbf{z})$ with respect to $z_i \in \mathbf{z}$ at $\mathbf{z}^o$ equals the sign of the first derivative of the one-way sensitivity function $\Pr(\mathbf{w}\,|\,\mathbf{u})(z_i)$ and therefore can be established efficiently Kjaerulff and van der Gaag (2000).

In a sliced sensitivity function Bolt and Renooij (2014) all parameters $\mathbf{z} = \{x, y_1, \ldots, y_{n-1}\}$ of an n-way sensitivity function are tied linearly. That is, all parameters $y_i \in \mathbf{z}$ are tied to x, by $y_i = \alpha_i \cdot x + \beta_i$. The constants $\alpha_1, \ldots \alpha_{n-1}$ determine a vector $\overrightarrow{\mathbf{a}} = (\alpha_1, \ldots, \alpha_{n-1})$ along which the parameters are changed. These constants can be chosen freely. The constants $\beta_1, \ldots \beta_{n-1}$ are chosen such that the function includes $\mathbf{z}^o$. The result is a function in x which takes the form of the fraction of two polynomials noted as

$$\Pr(\mathbf{w}\,|\,\mathbf{u})(x\|\mathbf{y})^{\overrightarrow{\mathbf{a}}} = \frac{c_0 + c_1 \cdot x^1 + \ldots + c_m \cdot x^m}{d_0 + d_1 \cdot x^1 + \ldots + d_m \cdot x^m}$$

where each x^k, $k = 1, \ldots, m$, is a polynomial term of degree k with $m \leq n$, and $\mathbf{y} = \{y_1, \ldots, y_{n-1}\}$. A sliced sensitivity function can be viewed as a linear section of an underlying n-way sensitivity function. Note that there are infinite many sliced sensitivity functions with the same underlying n-way function since $\overrightarrow{\mathbf{a}}$ can be chosen freely.

Also in a balanced sensitivity function Bolt and van der Gaag (2015) all parameters $\mathbf{z} = \{x, y_1, \ldots, y_{n-1}\}$ of the underlying n-way function are tied. Now the parameters y_i are tied to x based on the odds ratio of their original values and new values. The relation between y_i can be chosen to be positive or negative. The relation $x^o \cdot (1-x)/(1-x^o) \cdot x = y_i^o \cdot (1-y_i)/(1-y_i^o) \cdot y_i \Leftrightarrow y_i = \frac{x \cdot y^o \cdot (1-x^o)}{x \cdot (y^o - x^o) + x^o \cdot (1-y^o)}$, ties x and y_i positively in which case x and y simultaneously vary from 0 to 1. The relation $x^o \cdot (1-x)/(1-x^o) \cdot x = (1-y_i^o) \cdot y_i / y_i^o \cdot (1-y_i) \Leftrightarrow y_i = \frac{x^o \cdot y^o \cdot (x-1)}{y^o \cdot (x-x^o) + x \cdot (x^o - 1)}$, ties x and y_i negatively in which case y_i varies from 1 to 0 where x varies from 0 to 1. The result is again a function in x which takes the form of the fraction of two polynomials, noted as

$$\Pr(\mathbf{w}\,|\,\mathbf{u})(x\|\mathbf{y}^+, \mathbf{y}^-) = \frac{c_0 + c_1 \cdot x^1 + \ldots + c_m \cdot x^m}{d_0 + d_1 \cdot x^1 + \ldots + d_m \cdot x^m}$$

where each x^k, $k = 1, \ldots, m$, is a polynomial term of degree k with $m \leq n$ and where the parameters $y_i \in \mathbf{y}^+$ are tied positively to x, and the parameters $y_i \in \mathbf{y}^-$ are tied negatively to x, with $\mathbf{y}^+ \cup \mathbf{y}^- = \{y_1, \ldots, y_{n-1}\}$. A balanced sensitivity function is a curved section of the underlying n-way which has as main characteristic that all parameters simultaneous change in the entire interval $\langle 0, 1 \rangle$. Note that there are multiple balanced functions with the same underlying n-way function since each y_i can be tied either positively or negatively to x.

2.2 Distance Measures

A common measure for comparing two probability distributions $\Pr^o$ and $\Pr$ over the same set of variables $\mathbf{V}$ is the Kullback-Leibler (KL) divergence Kullback and Leibler (1951). This divergence is defined as $\mathrm{KL} = \sum_{\mathbf{v} \in \mathbf{V}} \Pr^o(\mathbf{v}) \cdot \ln \frac{\Pr^o(\mathbf{v})}{\Pr(\mathbf{v})}$. The KL-divergence is positive and equals zero if and only if $\Pr^o = \Pr$. Although the measure is asymmetric and therefore is not a proper distance measure it is often indicated as the KL-distance.

In Chan and Darwiche (2005) another measure for the distance between two distributions $\Pr^o$ and $\Pr$ over the same set of variables $\mathbf{V}$ was proposed, This distance CD is define as $\mathrm{CD} =$

$\ln \max_{\mathbf{v} \in \mathbf{V}} \frac{\Pr(\mathbf{v})}{\Pr^o(\mathbf{v})} - \ln \min_{\mathbf{v} \in \mathbf{V}} \frac{\Pr(\mathbf{v})}{\Pr^o(\mathbf{v})}$, where $0/0$ and ∞/∞ are defined to equal 1. The distance is positive, and equals zero if and only if $\Pr^o = \Pr$. In contrast to the KL-divergence, moreover, this measure is symmetric. It was shown that the KL-divergence and the CD-distance may result in a different outcome when used for judging whether a distribution $\Pr'$ or a distribution $\Pr''$ is closer to a distribution $\Pr$. [1]

3. Tuning Heuristics

In this section first the locally optimal heuristic Bolt and Renooij (2014) and the balanced heuristic Chan and Darwiche (2004); Bolt and van der Gaag (2015) are described. Then, a heuristic which combines the advantages of the two former heuristics is proposed. The essence of all three heuristics is that the parameters used for tuning are tied according a specific scheme. The changes induced in an output probability then can be described by a corresponding 'constrained' n-way sensitivity function. Note that a network can be tuned without the availability of the appropriate sensitivity function; the required parameter changes can be found by iterating the tied parameters towards a desired outcome. In Section 5 the heuristics are experimentally compared with respect to the KL and CD-distances they induce between original and tuned distributions in a small example network.

Definition 1 (Locally optimal heuristic) *Consider a set parameters $\mathbf{z} = \{x, y_1, \ldots, y_{n-1}\}$ of a Bayesian network $\mathcal{B}$. In a locally optimal tuning scheme, with respect to an output probability $\Pr(\mathbf{w} \mid \mathbf{u})$ of $\mathcal{B}$, all parameters $y_i \in \mathbf{z}$ are tied to x by:*

$$y_i = \frac{sy_i^o}{sx^o} \cdot (x - x^o) + y_i^o$$

where sx^o and sy_i^o are the values of the partial derivatives of $\Pr(\mathbf{w} \mid \mathbf{u})(\mathbf{z})$ with respect to x and y_i at $\mathbf{z}^o$, respectively.

In a locally optimal tuning heuristic, parameters y_i are tied linearly to x based on the values of the partial derivatives of the n-way sensitivity function $\Pr(\mathbf{w} \mid \mathbf{u})(\mathbf{z})$ at the original parameter values $\mathbf{z}^o$. Note that if sy_i^o/sx^o is positive, y_i and x, simultaneously increase/decrease and if sy_i^o/sx^o is negative, an increase of x results in an decrease of y_i and vice versa. Note furthermore that sx^o should be unequal zero, that is, a change of x should have an effect on $\Pr(\mathbf{w} \mid \mathbf{u})$. The change of an output probability given a locally optimal scheme is expressed by a sliced sensitivity function of which the vector $\overrightarrow{\mathbf{a}}$ equals the gradient of $\Pr(\mathbf{w} \mid \mathbf{u})(\mathbf{z})$ at $\mathbf{z}^o$. This implies that as long as the required changes are small, the changes will be a good approximation of the adjustments which satisfy the required constrained with a minimal total absolute change of the parameters. A sliced sensitivity function capturing the locally optimal scheme will be denoted by $\Pr(\mathbf{w} \mid \mathbf{u})(x \| y_1, \ldots, y_{n-1})^{\nabla}$.

Definition 2 (Balanced heuristic) *Consider a set parameters $\mathbf{z} = \{x, y_1, \ldots, y_{n-1}\}$ of a Bayesian network $\mathcal{B}$. In a balanced tuning scheme, with respect to an output probability $\Pr(\mathbf{w} \mid \mathbf{u})$ of $\mathcal{B}$, all parameters $y_i \in \mathbf{z}$ are tied to x by*

$$y_i = \begin{cases} \dfrac{x \cdot (x^o - 1) \cdot y_i^o}{x^o \cdot (x - 1 + y_i^o) - x \cdot y_i^o} & \text{if } y_i \updownarrow x \\ \dfrac{(x - 1) \cdot x^o \cdot y_i^o}{-x + x \cdot x^o + x \cdot y_i^o - x^o \cdot y_i^o} & \text{if } y_i \not\updownarrow x \end{cases}$$

1. In Chan and Darwiche (2004) it was proposed that in tuning an output probability, a balanced change of the parameters involved minimises the CD-distance. This, however, is not true as was recognised in Chan (2005).

The above relationships imply that if $y_i \updownarrow x$, we have that y_i and x simultaneously increase/decrease and that if $y_i \not\updownarrow x$, an increase of x results in an decrease of y_i and vice versa. Moreover, x and y_i vary simultaneously in the entire interval $\langle 0, 1\rangle$. The balanced sensitivity function $\Pr(\mathbf{w} \,|\, \mathbf{u})(x \| \mathbf{y}^+, \mathbf{y}^-)$, with $\mathbf{y}^+ = \{y_i \in \mathbf{z} \mid y_i \updownarrow x\}$ and $\mathbf{y}^- = \{y_i \in \mathbf{z} \mid y_i \not\updownarrow x\}$, capturing the balanced heuristic, will be noted as $\Pr(\mathbf{w} \,|\, \mathbf{u})(x \| \mathbf{y}^+, \mathbf{y}^-)^*$. Note that the condition $y_i \updownarrow x$ ($y_i \not\updownarrow x$) is fulfilled if the partial derivatives of $\Pr(\mathbf{w} \,|\, \mathbf{u})(\mathbf{z})$ with respect to y_i and x have the same (opposite) sign at $\mathbf{z}^o$. This implies that the heuristic results in the same qualitative relationships between x and y_i as the locally optimal heuristic. Now however, in contrast to the locally optimal heuristic, all parameters can vary in the entire interval $\langle 0, 1\rangle$. The range of $\Pr(\mathbf{w} \,|\, \mathbf{u})(x \| \mathbf{y}^+, \mathbf{y}^-)^*$ will therefore in general be larger than the range of $\Pr(\mathbf{w} \,|\, \mathbf{u})(x \| \mathbf{y})^\nabla$. Under the condition that all parameters involved have a guaranteed qualitative effect, the range of the balanced sensitivity function $\Pr(\mathbf{w} \,|\, \mathbf{u})(x \| \mathbf{y}^+, \mathbf{y}^-)^*$ will even be the same as the range of the underlying n-way sensitivity function $\Pr(\mathbf{w} \,|\, \mathbf{u})(x, \mathbf{y}^+, \mathbf{y}^-)$. Under this condition, the balanced heuristic thus covers the maximal tuning range. The function then will be noted as $\Pr(\mathbf{w} \,|\, \mathbf{u})(x \| \mathbf{y}^+, \mathbf{y}^-)^{\circledast}$. In Bolt et al. (to appear), Bayesian network parameters with such a guaranteed qualitative were identified.

The advantage of the locally optimal heuristic is that it induces locally the maximal effect on the tuned output; disadvantage however is that the heuristic will, in general, cover a smaller tuning range than the balanced heuristic. Disadvantage of the balanced heuristic on the other hand, is that this heuristic may be quite suboptimal for small network adaptations. Below now a tuning heuristic is defined which is locally optimal, yet covers the same tuning range as the balanced heuristic.

Definition 3 (Combined heuristic) *Consider a set parameters $\mathbf{z} = \{x, y_1, \ldots, y_{n-1}\}$ of a Bayesian network $\mathcal{B}$. In a combined tuning scheme, with respect to an output probability $\Pr(\mathbf{w} \,|\, \mathbf{u})$ of $\mathcal{B}$, all parameters $y_i \in \mathbf{z}$ are tied to x by:*

$$y_i = \begin{cases} \dfrac{x \cdot (sx^o \cdot y_i^o \cdot (y_i^o - 1) + sy_i^o \cdot (1 - x^o)) + sx^o \cdot y_i^o \cdot (1 - y_i^o) - sy_i^o \cdot x^o \cdot (1 - x^o)}{x \cdot (sx^o \cdot (y_i^o - 1) + sy_i^o \cdot (1 - x^o)) + sx^o \cdot (1 - y_i^o) - sy_i^o \cdot x^o \cdot (1 - x^o)} & \textit{if } y_i \updownarrow x,\ x \geq x^o \\ \dfrac{x \cdot sx^o \cdot {y_i^o}^2}{x \cdot (sx^o \cdot y_i^o - sy_i^o \cdot x^o) + sy_i^o \cdot {x^o}^2} & \textit{if } y_i \updownarrow x,\ x \leq x^o \\ \dfrac{x \cdot sx^o \cdot {y_i^o}^2 - sx^o \cdot {y_i^o}^2}{x \cdot (sy_i^o \cdot (1 - x^o) + sx^o \cdot y_i^o) + sy_i^o \cdot x^o \cdot (x^o - 1) - sx^o \cdot y_i^o} & \textit{if } y_i \not\updownarrow x,\ x \geq x^o \\ \dfrac{x \cdot (sx^o \cdot y_i^o \cdot (y_i^o - 1) - sy_i^o \cdot x^o) + sy_i^o \cdot {x^o}^2}{x \cdot (sx^o \cdot (y_i^o - 1) - sy_i^o \cdot x^o) + sy_i^o \cdot {x^o}^2} & \textit{if } y_i \not\updownarrow x,\ x \leq x^o \end{cases}$$

where $sx^o, sy_i^o, \ldots, sy_{n-1}^o$ are the values of the partial derivatives of $\Pr(\mathbf{w} \,|\, \mathbf{u})(\mathbf{z})$ with respect to respectively $x, y_1, \ldots, y_{n-1}$ at $\mathbf{z}^o$.

In the combined heuristic the relationship between y_i and x is constructed such that changes are locally optimal, *and* all parameters change simultaneously in the entire interval $\langle 0, 1\rangle$, tied either positively or negatively, depending the sign of their partial derivatives. The values of the first derivatives of the functions above equal sy_i^o / sx^o at x^o. This is also the value of the first derivate at x^o given relationship between y_i and x as defined for the locally optimal heuristic. The combined heuristic thus captures, just as the locally optimal heuristic a locally optimal change. Moreover the sign of the first derivatives of all the functions equals the sign of $sx^o \cdot sy_i^o$. If $y_i \updownarrow x$, the function thus is increasing and if $y_i \not\updownarrow x$ the function is decreasing. Furthermore for $x = 0$, we find $y_i = 0$ if $y_i \updownarrow x$ and $y_i = 1$ if $y_i \not\updownarrow x$ and for $x = 1$ we find $y_i = 1$ if $y_i \updownarrow x$ and $y_i = 0$ if $y_i \not\updownarrow x$. The function

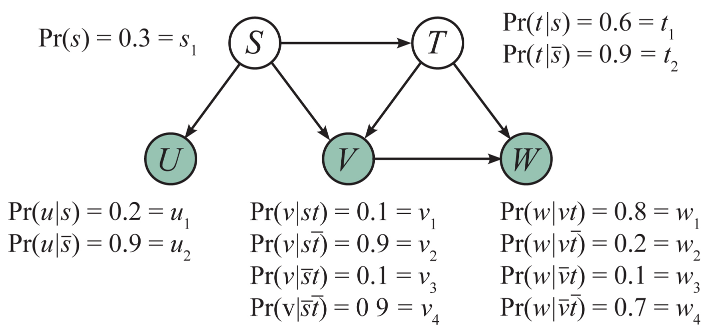

Figure 1: A small example network.

describing $\Pr(\mathbf{w}\,|\,\mathbf{u})$ given the change of the parameters $\mathbf{z}$ in the combined tuning heuristic will be indicated by $\Pr(\mathbf{w}\,|\,\mathbf{u})(x\|\mathbf{y}^+, \mathbf{y}^-)^{\star,\nabla}$ in which $\mathbf{y}^+ = \{y_i \in \mathbf{z} \mid y_i \updownarrow x\}$ and $\mathbf{y}^- = \{y_i \in \mathbf{z} \mid y_i \not\updownarrow x\}$. If the heuristic just involves parameters with a guaranteed qualitative effect, the range of this function will be the same as the range of the underlying n-way function. The sensitivity function will then be indicated by $\Pr(\mathbf{w}\,|\,\mathbf{u})(x\|\mathbf{y}^+, \mathbf{y}^-)^{\circledast,\nabla}$.

4. Example

Consider the example network from Figure 1. In the current network it is found that $\Pr(st\,|\,uvw) = 0.045$. Now suppose that $\Pr(st\,|\,uvw)$ has to be tuned to 0.25 by adapting the parameters $v_1 = \Pr(v\,|\,st)$ and $w_2 = \Pr(w\,|\,v\bar{t})$. The 2-way sensitivity function of $\Pr(st\,|\,uvw)$ in v_1 and w_2 is

$$\Pr(st\,|\,uvw)(v_1, w_2) = \frac{2.88 \cdot v_1}{2.88 \cdot v_1 + 7.83 \cdot w_2 + 4.54}$$

The extremes of this function are $\Pr(st\,|\,uvw)(0, 1) = 0$ and $\Pr(st\,|\,uvw)(1, 0) = 0.388$. The different tuning heuristics from Section 3 now define different relationships between v_1 and w_2. For a locally optimal heuristic is found

$$w_2 = -0.128 \cdot v_1 + 0.213$$

As can be established using the results in Bolt et al. (to appear), v_1 has a guaranteed positive effect on $\Pr(st\,|\,uvw)$ and w_2 has a guaranteed negative effect. Since the parameters v_1 and w_2 have opposite effects on the output probability, that is, $v_1 \not\updownarrow w_2$, for the balanced heuristic is found

$$w_2 = \frac{0.2 \cdot v_1 - 0.2}{-7 \cdot v_1 - 0.2}$$

And for the combined heuristic is found

$$w_2 = \begin{cases} \dfrac{1.721 - 1.722 \cdot v_1}{8.111 - 3.637 \cdot v_1} & \text{if } v_1 \geq v_1^o \\ \dfrac{0.055 + 6.334 \cdot v_1}{0.055 + 33.88 \cdot v_1} & \text{if } v_1 \leq v_1^o \end{cases}$$

The different relationships between v_1 and w_2 are depicted in Figure 2. The figure shows that, for the combined and the locally optimal heuristic, the functions relating v_1 and w_2 have the same

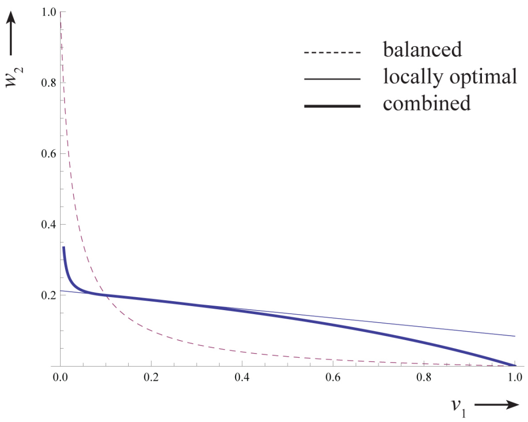

Figure 2: A balanced, a locally optimal and a combined relationship between the parameters v_1 and w_2 given the example network of Figure 1 and the output probability $\Pr(st \mid uvw)$.

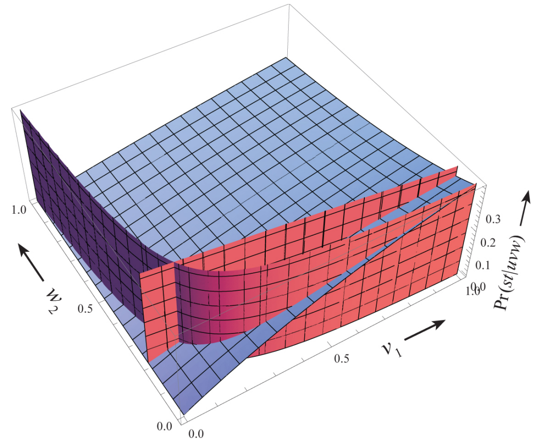

Figure 3: The 2-way sensitivity function $\Pr(st \mid uvw)(v_1, w_2)$, the curved surface determining the function $\Pr(st \mid uvw)(v_1 \| w_2^-)^{\circledast}$ and the plane determining the function $\Pr(st \mid uvw)(v_1 \| w_2)^{\nabla}$.

slope at $(v_1^o, w_2^o) = (0.1, 0.2)$. Given these heuristics the simultaneous changes of v_1 and w_2 are locally optimal. The figure moreover shows that for the combined and the balanced heuristic, the parameters v_1 and w_2 both cover the entire interval $\langle 0, 1 \rangle$.

Figure 3 shows the 2-way function $\Pr(st \mid uvw)(v_1, w_2)$ and shows in addition the two surfaces which meet the constraints established for the relationship between v_1 and w_2 given the locally optimal and the balanced heuristic, determining the functions $\Pr(st \mid uvw)(v_1 \| w_2)^{\nabla}$ and $\Pr(st \mid uvw)(v_1 \| w_2^-)^{\circledast}$. (The surface determining $\Pr(st \mid uvw)(v_1 \| w_2^-)^{\circledast,\nabla}$ is not depicted.)
The constrained functions capturing the different heuristics can be derived from the two-way sensitivity function using the relationships between v_1 and w_2 given above, These functions are

$$\Pr(st \mid uvw)(v_1 \| w_2)^{\nabla} = \frac{2.880 \cdot v_1}{6.202 + 1.875 \cdot v_1}$$

$$\Pr(st \mid uvw)(v_1 \| w_2^-)^{\circledast} = \frac{0.576 \cdot v_1 + 20.16 \cdot v_1^2}{2.474 + 30.79 \cdot v_1 + 20.16 \cdot v_1^2}$$

and

$$\Pr(st \mid uvw)(v_1 \| w_2^-)^{\circledast,\nabla} = \begin{cases} \dfrac{-2.230 \cdot v_1 + v_1^2}{-4.799 + 0.632 \cdot v_1 + v_1^2} & \text{if } v_1 \geq v_1^o \\ \dfrac{-0.163 \cdot v_1 - 100 \cdot v_1^2}{0.700 - 208.5 \cdot v_1 - 100 \cdot v_1^2} & \text{if } v_1 \leq v_1^o \end{cases}$$

The desired output $\Pr(st \mid uvw) = 0.25$ is found for $v_1 = 0.643$ (and $w_2 = 0.130$), for $v_1 = 0.546$ (and $w_2 = 0.0152$) and for $v_1 = 0.625$ (and $w_2 = 0.111$), respectively.

The distances between the probability distributions represented by the original and the adapted network given the different heuristics are

	locally optimal	balanced	combined
CD-distance	2.79	3.88	2.71
KL-divergence	0.120	0.127	0.115

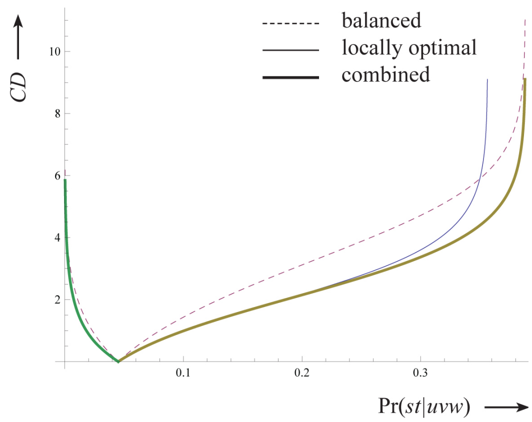

Figure 4: The CD-distance between the original and the tuned network from Figure 1 as a function of $\Pr(st\,|\,uvw)$ given a balanced, a locally optimal and a combined tuning of this probability with v_1 and w_2.

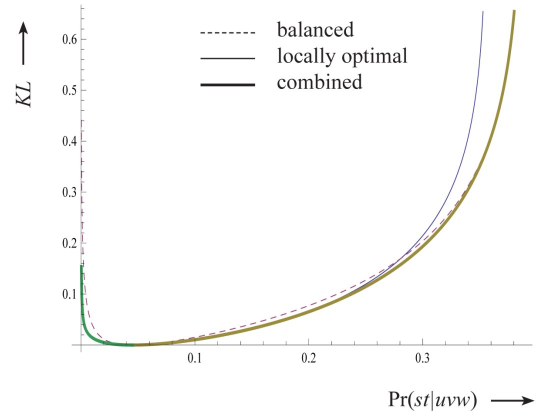

Figure 5: The KL-distance between the original and the tuned network from Figure 1 as a function of $\Pr(st\,|\,uvw)$ given a balanced, a locally optimal and a combined tuning of this probability with v_1 and w_2.

For tuning $\Pr(st\,|\,uvw)$ to 0.25, both with respect to minimising the CD-distance as with respect to minimising the KL-divergence, the combined heuristic thus is preferred.

To illustrate the relation between the CD-distance and the required value of the tuned probability, Figure 4 shows the CD-distance between the original and the tuned network as a function of $\Pr(st\,|\,uvw)$ given locally optimal, balanced and combined tuning with v_1 and w_2. The figure shows that the combined heuristic performs at least as well as both other heuristics with respect to minimising the CD-distance. Moreover the figure shows that the balanced and the combined heuristic can be used for the entire tuning range $\langle 0, 0.388\rangle$ whereas the locally optimal heuristic cannot be applied for new values of $\Pr(st\,|\,uvw)$ above 0.357.

To conclude the example, Figure 5 shows the KL-divergence as a function of the new value of $\Pr(st\,|\,uvw)$ for linear, balanced and combined tuning with v_1 and w_2. Although less pronounced, the figure shows similar tendencies as Figure 4.

5. Experiments

To investigate whether the combined tuning heuristic proposed in Section 3 might be a suitable alternative in tuning a network with multiple parameters, preliminary experiments were performed with the network from Figure 1. For this small network the CD-distance and the KL-divergence could be established exactly. The locally optimal (L), balanced (B) and combined (C) heuristic were compared with respect to the CD-distance and the KL-divergence between original and the tuned network when tuning the output value $\Pr(st\,|\,uvw)$. First, 100 sets of parameter values in the interval $\langle 0, 1\rangle$ were of randomly generated from a uniform distribution.[2] For each of the 100 resulting networks, the original probability $\Pr^o(st\,|\,uvw)$ was, using the proposed heuristics, tuned to new values $\Pr(st\,|\,uvw)$ such that $\ln\left(\frac{(1-\Pr^o(st\,|\,uvw))\cdot\Pr(st\,|\,uvw)}{\Pr^o(st\,|\,uvw)\cdot(1-\Pr((st\,|\,uvw))}\right)$ equaled -2, -1, -0.5, 0.5, 1

2. Since the example network is binary, for each local distribution one parameter value $value$ was generated; the other parameter of the same distribution was set to $1 - value$.

and 2. An output value of 0.2, for example, would be tuned to 0.033, 0.084, 0.132, 0.292, 0.405 and 0.649 an output value of 0.5 would be tuned to 0.119, 0.269, 0.378, 0.622, 0.731 and 0.881. The heuristics were compared for the following sets of tuning parameters

$$\begin{aligned} \text{set1} &= \{v_1, u_2, w_2\} \\ \text{set2} &= \{v_1, v_2, v_3, v_4\} \\ \text{set3} &= \{v_1, v_2, u_1, u_2, w_1, w_2\} \end{aligned}$$

These sets all include v_1. For $v_1 = 0$ is found that $\Pr(st\,|\,uvw) = 0$ which implies that including v_1 guarantees that tuning to lower values is always feasible when using the balanced or the combined heuristic. The experiments were repeated with 100 sets of parameter assignments generated from a distribution biased towards 0 and 1.[3]

	'uniform'			'biased'		
	set 1	set 2	set 3	set 1	set 2	set 3
$\ln(OR) = -2$	31	35	13	40	43	10
$\ln(OR) = -1$	19	14	6	30	34	5
$\ln(OR) = -0.5$	9	7	3	21	22	3
$\ln(OR) = 0.5$	9	9	6	22	35	9
$\ln(OR) = 1$	24	32	13	32	45	18
$\ln(OR) = 2$	50	68	35	53	58	35

Table 1: The number out of 100 networks with the graph from Figure 1 with parameters generated randomly from a uniform distribution cq from a distribution biased towards 0 and 1, in which the output probability $\Pr(st\,|\,uvw)$ could not be tuned to the desired value.

The results are presented in Tables 1 and 2. Table 1 shows the number of networks for which the output could not be tuned to the desired value using the locally optimal heuristic. As could be expected, the larger the difference between the original and the desired output probability, the larger the number of networks which could not be tuned. Moreover, the 'biased parameterisation' resulted in substantially more networks that could not be tuned than the 'uniform parameterisation' when tuning with set 1 or set 2. This result may be explained by the 'biased parameterisation' resulting in a more diverging impact of the individual parameters on the output. Given the small tuning sets, then there may be just parameters with a small impact included. For the other two tuning heuristics was observed that in all cases the desired output probability could be reached.

The heuristics were then pairwise compared with respect to the distances between the original and tuned networks. Table 2 indicates which heuristic most often resulted in the smallest CD/KL-distance for the different tuning values, different tuning sets and different parameterisations. When an output probability could not be tuned to the desired value, the distance was supposed to be ∞. With t a 50/50 result is indicated. The underlying numbers are given in the appendix.

3. The biased parameters were sampled from an U-quadratic distribution, with lower limit 0 and upper limit 1.

	L vs B			L vs C			B vs C		
	set 1	set 2	set 3	set 1	set 2	set 3	set 1	set 2	set 3
CD-distance, ‘uniform parameterisation’									
$\ln(OR) = -2$	L	L	L	C	C	L	C	C	C
$\ln(OR) = -1$	L	L	L	C	C	L	C	C	C
$\ln(OR) = -0.5$	L	L	L	C	C	L	C	C	C
$\ln(OR) = 0.5$	L	B	L	C	C	C	C	C	C
$\ln(OR) = 1$	L	B	L	C	C	C	C	C	C
$\ln(OR) = 2$	B	B	B	C	C	C	C	B	C
CD-distance, ‘biased parameterisation’									
$\ln(OR) = -2$	B	B	L	C	C	L	C	C	C
$\ln(OR) = -1$	L	L	L	C	C	L	C	C	C
$\ln(OR) = -0.5$	L	L	L	C	C	L	C	C	C
$\ln(OR) = 0.5$	L	B	L	C	C	C	C	C	C
$\ln(OR) = 1$	L	B	L	C	C	C	C	C	C
$\ln(OR) = 2$	B	B	L	C	C	C	C	C	C
KL-divergence, ‘uniform parameterisation’									
$\ln(OR) = -2$	L	L	L	C	t	t	C	C	C
$\ln(OR) = -1$	L	L	L	C	L	C	C	C	C
$\ln(OR) = -0.5$	L	L	L	C	L	t	C	C	C
$\ln(OR) = 0.5$	L	L	L	C	C	C	C	C	C
$\ln(OR) = 1$	B	B	L	C	C	C	C	C	C
$\ln(OR) = 2$	B	B	B	C	C	C	C	C	C
KL-divergence, ‘biased parameterisation’									
$\ln(OR) = -2$	B	B	L	C	C	L	C	C	C
$\ln(OR) = -1$	L	B	L	C	C	L	C	C	C
$\ln(OR) = -0.5$	L	B	L	C	C	L	C	C	C
$\ln(OR) = 0.5$	L	B	L	C	C	C	C	C	C
$\ln(OR) = 1$	L	B	L	C	C	C	C	C	C
$\ln(OR) = 2$	B	B	L	C	C	C	C	C	C

Table 2: Pairwise comparison of the balanced (B), locally optimal (L) and combined (C) tuning heuristic, for six different tuning values, three different tuning sets and two types of parameterisation. The indicated heuristic resulted in the majority of 100 networks in the smallest distance between the original and the tuned distribution.

The results are quite similar for as well the CD/KL-distance as for two types of parameterisation. In comparing the locally optimal and the balanced heuristic it is observed that, using set 1 or set 2, the locally optimal heuristic more often induced the smallest distance for tuning to values closer to the original probability, whereas the balanced heuristic more often induced the smallest distance for the more extreme tuning values, especially in tuning to higher values. Using set 3, tuning with the locally optimal heuristic was most favourable in almost all cases. In comparing the locally optimal with the combined heuristic, it is observed that the combined heuristic most often resulted in the smallest distance/divergence in most cases. Tuning with the locally optimal heuristic was only most favourable for tuning to lower values given the larg(er)(est) tuning sets 2 and 3. In comparing the balanced and the combined heuristic, to conclude, it is observed that for almost all cases the combined heuristic resulted most often in the smallest CD/KL-distance.

All in all, the preliminary experiments with the combined tuning heuristic show that this heuristic is a promising alternative in tuning networks with multiple parameters with respect to minimising the CD/KL-distance between the distributions represented by the original and the tuned networks.

6. Conclusions

One of the issues in tuning an output probability of a Bayesian network by changing multiple parameters is the relative amount of change of the individual parameters; a combination of parameter changes which disturbs the original network as little as possible may be preferred. An existing 'locally optimal' heuristic induces a locally optimal combination of parameter changes; another existing 'balanced' heuristic has as advantage that it covers, in general, a larger range of attainable values than the locally optimal heuristic. In this paper a new tuning heuristic was proposed that is locally optimal, yet covers the same tuning range as the balanced heuristic. In preliminary experiments with a small example network, this combined heuristic often outperformed the two existing heuristics with respect to minimising the CD/KL-distance between the original and the tuned network in tuning one of its output values. The combined heuristic thus is a promising alternative for network tuning with multiple parameters.

More research is required to refine the results of this paper. A subject for future research, is the effect of the inclusion of parameters without a guaranteed qualitative effect in the tuning set. In that case, although the tuning range will in general still be larger than the tuning range given the locally optimal heuristic, also the balanced and the combined heuristic are not guaranteed to cover the maximal tuning range any more. Also the effect of the choice of the tuning parameters needs more research. Selecting just parameters with an individual high local influence will favour the locally optimal heuristics; other criteria for parameter selection, for example, the selection of the parameters which were based on just a few data, may have a different effect on the performance of the heuristics. Yet another interesting subject for future research is the relation between the solutions found by using the proposed heuristics and the solution that minimises the distance between the distributions represented by the original and the tuned network.

Acknowledgement

This work was supported by the Netherlands Organisation for Scientific Research.

References

J. Bolt and S. Renooij. Local sensitivity of bayesian networks to multiple parameter shifts. In L. van der Gaag and A. Feelders, editors, *Proceedings of the 7th European Workshop on Probabilistic Graphical Models*, pages 65–80, 2014.

J. Bolt and L. van der Gaag. Balanced tuning of multi-dimensional bayesian network classifiers. In S. Destrecke and T. Denoeux, editors, *Proceedings of the 13th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 210–220, 2015.

J. Bolt, J. D. Bock, and S. Renooij. Exploiting bayesian network sensitivity functions for inference in credal networks. In *Proceedings of the 22nd European Conference on Artificial Intelligence*, to appear.

H. Chan. *Sensitivity analysis of probabilistic graphical models*. PhD thesis, 2005.

H. Chan and A. Darwiche. Sensitivity analysis in bayesian networks: From single to multiple parameters. In M. Chickering and J. Halpern, editors, *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 67–75. AUAI Press, Arlington, VA, 2004.

H. Chan and A. Darwiche. A distance measure for bounding probabilistic belief change. *International Journal of Approximate Reasoning*, 38:149–174, 2005.

V. Coupé and L. van der Gaag. Properties of sensitivity analysis of bayesian belief networks. *Annals of Mathematics and Artificial Intelligence*, 36:323–356, 2002.

U. Kjaerulff and L. van der Gaag. Making sensitivity analysis computationally efficient. In C. Boutilier and M. Goldszmidt, editors, *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 317–325, San Francisco, 2000. Morgan Kaufmann.

S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22: 79–86, 1951.

Appendix

	CD-distance, 'uniform parameterisation'								
	$L-B>0$			$L-C>0$			$B-C>0$		
	set 1	set 2	set 3	set 1	set 2	set 3	set 1	set 2	set 3
$\ln(OR)=-2$	46	43	23	62	84	34	75	79	78
$\ln(OR)=-1$	33	38	23	62	82	33	79	78	83
$\ln(OR)=-0.5$	28	34	22	60	83	38	77	73	80
$\ln(OR)=0.5$	32	52	25	69	94	70	77	54	77
$\ln(OR)=1$	39	63	31	70	99	73	80	55	76
$\ln(OR)=2$	64	88	57	83	99	84	69	33	71

	CD-distance, 'biased parameterisation'								
	$L-B>0$			$L-C>0$			$B-C>0$		
	set 1	set 2	set 3	set 1	set 2	set 3	set 1	set 2	set 3
$\ln(OR)=-2$	53	76	74	59	76	64	17	31	84
$\ln(OR)=-1$	40	72	78	46	77	68	13	31	84
$\ln(OR)=-0.5$	37	69	74	48	74	64	17	32	84
$\ln(OR)=0.5$	39	72	69	57	95	52	25	64	76
$\ln(OR)=1$	45	76	70	57	95	54	31	64	74
$\ln(OR)=2$	60	82	67	71	99	55	44	71	71

	KL-divergence 'uniform parameterisation'								
	$L-B>0$			$L-C>0$			$B-C>0$		
	set 1	set 2	set 3	set 1	set 2	set 3	set 1	set 2	set 3
$\ln(OR)=-2$	49	48	25	51	50	50	66	87	87
$\ln(OR)=-1$	41	36	21	54	49	51	67	84	86
$\ln(OR)=-0.5$	37	31	22	52	47	50	69	81	83
$\ln(OR)=0.5$	48	48	25	62	66	65	60	67	81
$\ln(OR)=1$	51	60	36	66	77	69	60	58	81
$\ln(OR)=2$	65	80	53	79	90	79	58	53	80

	KL-divergence, 'biased parameterisation'								
	$L-B>0$			$L-C>0$			$B-C>0$		
	set 1	set 2	set 3	set 1	set 2	set 3	set 1	set 2	set 3
$\ln(OR)=-2$	54	65	69	57	61	73	21	38	84
$\ln(OR)=-1$	46	61	71	57	59	70	18	42	87
$\ln(OR)=-0.5$	42	60	68	51	58	68	16	44	86
$\ln(OR)=0.5$	36	62	76	56	62	62	26	60	80
$\ln(OR)=1$	46	65	75	59	68	65	31	65	81
$\ln(OR)=2$	60	76	76	71	82	64	41	70	79

Pairwise comparison of the balanced (B), locally optimal (L) and combined (C) tuning heuristic, for six different tuning values, three different tuning sets and two types of parameterisation. The number indicates the number of times out of 100 that distance concerned was larger than zero.