Appeared in: International Journal of Approximate Reasoning, Vol. 52, No. 5, July 2011, pp. 641--657.

SCHOOL OF BUSINESS WORKING PAPER NO. 321

Inference in Hybrid Bayesian Networks Using Mixtures of Polynomials

Prakash P. Shenoy

and

James C. West

September 16, 2010†

University of Kansas School of Business
1300 Sunnyside Ave, Summerfield Hall
Lawrence, KS 66045-7601 USA
<pshenoy@ku.edu>, <cully@ku.edu>

† Accepted for publication in the *International Journal of Approximate Reasoning*.

Table of Contents

Inference in Hybrid Bayesian Networks Using Mixtures of Polynomials

Prakash P. Shenoy and *James C. West*
University of Kansas School of Business, 1300 Sunnyside Ave.,
Summerfield Hall, Lawrence, KS 66045-7601 USA
<pshenoy@ku.edu>, <cully@ku.edu>

Abstract

The main goal of this paper is to describe inference in hybrid Bayesian networks (BNs) using mixture of polynomials (MOP) approximations of probability density functions (PDFs). Hybrid BNs contain a mix of discrete, continuous, and conditionally deterministic random variables. The conditionals for continuous variables are typically described by conditional PDFs. A major hurdle in making inference in hybrid BNs is marginalization of continuous variables, which involves integrating combinations of conditional PDFs. In this paper, we suggest the use of MOP approximations of PDFs, which are similar in spirit to using mixtures of truncated exponentials (MTEs) approximations. MOP functions can be easily integrated, and are closed under combination and marginalization. This enables us to propagate MOP potentials in the extended Shenoy-Shafer architecture for inference in hybrid BNs that can include deterministic variables. MOP approximations have several advantages over MTE approximations of PDFs. They are easier to find, even for multi-dimensional conditional PDFs, and are applicable for a larger class of deterministic functions in hybrid BNs.

1 Introduction

Bayesian networks (BNs) and influence diagrams (IDs) were invented in the mid 80s (see e.g., [Pearl 1986], [Howard and Matheson 1984]) to represent and reason with large multivariate discrete probability models and decision problems, respectively. Several efficient algorithms exist to compute exact marginals of posterior distributions for discrete BNs (see e.g., [Lauritzen and Spiegelhalter 1988], [Shenoy and Shafer 1990], and [Jensen *et al.* 1990]) and to solve discrete IDs exactly (see e.g., [Olmsted 1983], [Shachter 1986], [Shenoy 1992], and [Jensen *et al*. 1994]).

Hybrid Bayesian networks contain a mix of discrete and continuous variables. A continuous variable is said to be *deterministic* if its conditional distributions have zero variances. The conditional distributions of deterministic variables are typically described by equations that describe the deterministic variable as a function of its continuous parents. Deterministic variables

pose a problem in inference since the joint density of all continuous variables does not exist. Shenoy and West [2009a] describe an extension of the Shenoy-Shafer architecture [Shenoy and Shafer 1990] to enable inference in hybrid BNs with deterministic variables.

The state of the art exact algorithm for mixtures of Gaussians hybrid BNs is the Lauritzen-Jensen [2001] algorithm implemented with Madsen's [2008] lazy propagation technique. This requires the conditional PDFs of continuous variables to be conditional linear Gaussians (CLGs), and that discrete variables do not have continuous parents. Marginals of multivariate normal distributions can be found easily without the need for integration. The disadvantages are that in the inference process, continuous variables have to be marginalized before discrete ones. In some problems, this restriction can lead to large cliques [Lerner and Parr 2001].

If a BN has discrete variables with continuous parents, Murphy [1999] uses a variational approach to approximate the product of the potentials associated with a discrete variable and its parents with a CLG distribution. Lerner [2002] uses a numerical integration technique called Gaussian quadrature to approximate non-CLG distributions with CLG distributions, and this same technique can be used to approximate the product of potentials associated with a discrete variable and its continuous parents. Murphy's and Lerner's approach is then embedded in the Lauritzen and Jensen [2001] algorithm to solve the resulting mixtures of Gaussians BN.

Shenoy [2006] proposes approximating non-CLG distributions by mixtures of Gaussians using a nonlinear optimization technique, and using arc reversals to ensure discrete variables do not have continuous parents. The resulting mixture of Gaussians BN is then solved using the Lauritzen and Jensen [2001] algorithm.

Moral *et al*. [2001] proposes approximating PDFs by mixtures of truncated exponentials (MTEs), which are easy to integrate in closed form. Since the family of mixtures of truncated exponentials is closed under combination and marginalization, the Shenoy-Shafer [1990] architecture can be used to solve an MTE BN. Cobb and Shenoy [2006] and Cobb *et al*. [2006] propose using a non-linear optimization technique for finding MTE approximations for several commonly used one-dimensional distributions. Cobb and Shenoy [2005a, b] extend this approach to BNs with linear and non-linear deterministic variables. In the latter case, they approximate non-linear deterministic functions by piecewise linear ones. Rumi and Salmeron [2007] describe approximate probability propagation with MTE approximations that have only two exponential terms in each piece. Romero *et al*. [2006] describe learning MTE potentials from data, and Langseth *et al*. [2010] investigate the use of MTE approximations where the coefficients are restricted to integers.

In this paper, we propose using mixture of polynomials (MOP) approximations of PDFs. Mixtures of polynomials are widely used in many domains including computer graphics, font design, approximation theory, and numerical analysis. They were first studied by Schoenberg

[1946]. When the MOP functions are continuous, they are referred to as *polynomial splines* [de Boor 2001, Schumaker 2007]. The use of splines to approximate PDFs was initially suggested by Curds [1997]. For our purposes, continuity is not an essential requirement, and we will restrict our analysis to piecewise polynomial approximations of PDFs.

Using MOP is similar in spirit to using MTEs. MOP functions can be easily integrated, and they are closed under combination and marginalization. Thus, the extended Shenoy-Shafer architecture [Shenoy and West 2009a] can be used to make inferences in BN with deterministic variables. However, there are several advantages of MOP functions over MTEs.

First, we can find MOP approximations of differentiable PDFs easily by using the Taylor series approximations. Finding MTE approximations as suggested by Cobb *et al*. [2006] necessitates solving non-linear optimization problems, which is not as easy a task as it involves navigating among local optimal solutions.

Second, for the case of conditional PDFs with several parents, finding a good MTE approximation can be extremely difficult as it involves solving a non-linear optimization problem in a high-dimensional space for each piece. The Taylor series expansion can also be used for finding MOP approximations of conditional PDFs. In Section 2, we describe a MOP approximation for a 2-dimensional CLG distribution.

Third, if a hybrid BN contains deterministic functions, then the MTE approach can be used directly only for linear deterministic functions. By directly, we mean without approximating a non-linear deterministic function by a piecewise linear one. This is because the MTE functions are not closed under transformations needed for non-linear deterministic functions. MOP functions are closed under a larger family of deterministic functions including linear functions and quotients. This enables propagation in a bigger family of hybrid BNs than is possible using MTEs.

An outline of the remainder of the paper is as follows. In Section 2, we define MOP functions and describe how one can find MOP approximations with illustration for the univariate normal distribution, chi-square distribution, and for a two-dimensional CLG distribution. In Section 3, we sketch the extended Shenoy-Shafer architecture for inference in hybrid BNs with deterministic variables. In Section 4, we solve three small examples designed to demonstrate the feasibility of using MOP approximations with non-linear deterministic functions. Finally, in Section 5, we end with a summary and discussion of some of the challenges associated with MOP approximations.

2 Mixtures of Polynomials Approximations

In this section, we describe MOP functions and some methods for finding MOP approximations of PDFs. We illustrate our method for the normal distribution, the chi-square distribution, and the CLG distribution in two dimensions.

A one-dimensional function $f\colon \mathbb{R} \to \mathbb{R}$ is said to be a *mixture of polynomials* (MOP) function if it is a piecewise function of the form:

$$\begin{aligned} f(x) &= a_{0i} + a_{1i}\, x + a_{2i}\, x^2 + \ldots + a_{ni}\, x^n, && \text{for } x \in A_i,\ i = 1, \ldots, k, \text{ and} \\ &= 0, && \text{otherwise,} \end{aligned} \tag{2.1}$$

where $A_1, \ldots, A_k$ are disjoint intervals in $\mathbb{R}$ that do not depend on x, and $a_{0i}, \ldots, a_{ni}$ are constants for all i. We will say that f is a *k-piece* (ignoring the 0 piece), and *n-degree* (assuming $a_{ni} \neq 0$ for some i) MOP function.

The main motivation for defining MOP functions is that such functions are easy to integrate in closed form, and that they are closed under multiplication and integration. They are also closed under differentiation and addition.

An m-dimensional function $f\colon \mathbb{R}^m \to \mathbb{R}$ is said to be a MOP function if

$$f(x_1, \ldots, x_m) = f_1(x_1) \cdot f_2(x_2) \cdot \ldots \cdot f_m(x_m), \tag{2.2}$$

where each $f_i(x_i)$ is a one-dimensional MOP function as defined in (2.1). If $f_i(x_i)$ is a k_i-piece and n_i-degree, then f is $k_1 \cdot \ldots \cdot k_m$ piece, degree $n_1 + \ldots + n_m$ MOP function. Therefore, it is important to keep the number of pieces and degrees to a minimum.

Example 2.1. Consider the univariate standard normal PDF $\varphi(z) = (1/\sqrt{2\pi})\, e^{-z^2/2}$. A 1-piece, 28-degree, MOP approximation $\varphi_{1p}(z)$ of $\varphi(z)$ in the interval $(-3, 3)$ is as follows:

$$\begin{aligned} \varphi_{1p}(z) &= c^{-1}(1 - z^2/2 + z^4/8 - z^6/48 + z^8/384 + \ldots + z^{28}/1{,}428{,}329{,}123{,}020{,}800), && \text{if } -3 < z < 3, \\ &= 0, && \text{otherwise,} \end{aligned}$$

where $c^{-1} \approx 0.4$. This MOP approximation was found using the Taylor series expansion of $e^{-z^2/2}$ at $z = 0$, to degree 28, restricting it to the region $(-3, 3)$, verifying that $\varphi_{1p}(z) \geq 0$ in the region $(-3, 3)$, and normalizing it with constant c so that $\int \varphi_{1p}(z) dz = 1$ (whenever the limits of integration of an integral are not specified, the entire range $(-\infty, \infty)$ is to be understood). We will denote these operations by writing:

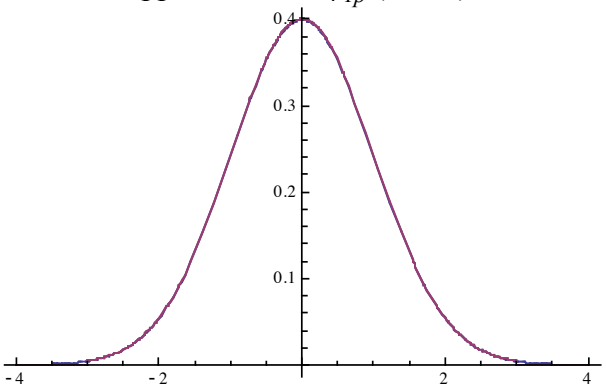
$$\begin{aligned} \varphi_{1p}(z) &= \mathit{TSeries}[e^{-z^2/2}, z = 0, d = 28] && \text{if } -3 < z < 3 \\ &= 0, && \text{otherwise} \end{aligned} \tag{2.3}$$

We can verify that $\varphi_{1p}(z) \geq 0$ as follows. First, we plot the un-normalized MOP approximation, denoted by, say, $\varphi_u(z)$. From the graph, we identify approximately the regions where $\varphi_u(z)$ could possibly be negative. Then starting from a point in each of these regions, we compute the local minimum of $\varphi_u(z)$ using, e.g., gradient descent. Since MOP functions are easily differentiable, the gradients can be easily found. If $\varphi_u(z) \geq 0$ at all the local minimums, then we have verified that $\varphi_{1p}(z) \geq 0$. If $\varphi_u(z) < 0$ at a local minimum, then we need to either increase the degree of the polynomial approximation, or increase the number of pieces, or both.

We have some very small coefficients in the MOP approximation. Rounding these off to a certain number of decimal places could cause numerical instability. Therefore, it is important to keep the coefficients in their rational form. A similar idea is proposed by Langseth *et al*. [2010] where they restrict the coefficients in MTE approximations to integers for reasons of stability.

A graph of the MOP approximation φ_{1p} overlaid on the actual PDF φ is shown in Figure 1 and it shows that there are not many differences between the two functions in the interval $(-3, 3)$. The main difference is that φ_{1p} is restricted to $(-3, 3)$, whereas φ is not. The mean of φ_{1p} is 0, and its variance ≈ 0.976. Most of the error in the variance is due to the restriction of the distribution to the interval $(-3, 3)$. If we restrict the standard normal density function to the interval $(-3, 3)$, renormalize it so that it is a PDF, then its variance ≈ 0.973.

Figure 1. The PDF of the standard normal distribution (in blue) and its polynomial approximation φ_{1p} (in red).

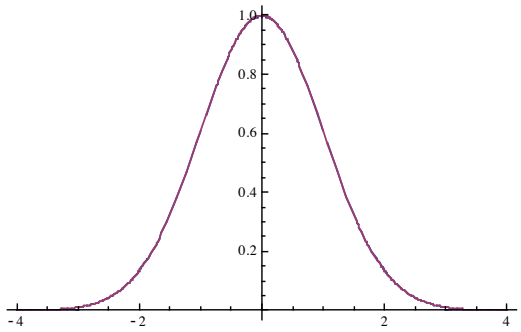


In some examples, working with a 28-degree polynomial may not be tractable. In this case, we can include more pieces to reduce the degree of the polynomial. For example, a 6-piece, 3-degree MOP approximation of $\varphi(z)$ is as follows:

$$
\begin{aligned}
\varphi_{6p}(z) &= TSeries[e^{-z^2/2},\ z = -5/2, d = 3] && \text{if } -3 \le z < -2,\\
&= TSeries[e^{-z^2/2},\ z = -3/2, d = 3] && \text{if } -2 \le z < -1,\\
&= TSeries[e^{-z^2/2},\ z = -1/2, d = 3] && \text{if } -1 \le z < 0,\\
&= TSeries[e^{-z^2/2},\ z = 1/2,\ \ d = 3] && \text{if } 0 \le z < 1,\\
&= TSeries[e^{-z^2/2},\ z = 3/2,\ \ d = 3] && \text{if } 1 \le z < 2,\\
&= TSeries[e^{-z^2/2},\ z = 5/2,\ \ d = 3] && \text{if } 2 \le z \le 3,\\
&= 0, && \text{otherwise.}
\end{aligned} \tag{2.4}
$$

Notice that φ_{6p} is discontinuous at the end points of the intervals. A graph of φ_{6p} overlaid on the graph of φ is shown in Figure 2. Also, $E(\varphi_{6p}) = 0$, and $V(\varphi_{6p}) \approx 0.974$. The variance of φ_{6p} is closer to the variance of the truncated normal (≈ 0.973) than φ_{1p}.

Figure 2. A graph of φ_{6p} (in blue) overlaid on the graph of φ (in red)



In some examples, for reasons of precision, we may wish to work with a larger interval than (−3, 3) for the standard normal. For example, an 8-piece, 4-degree MOP approximation of φ in the interval (−4, 4) is as follows:

$$
\begin{aligned}
\varphi_{8p}(z) &= TSeries[e^{-z^2/2},\ z = -7/2, d = 4] && \text{if } -4 < z < -3,\\
&= TSeries[e^{-z^2/2},\ z = -5/2, d = 3] && \text{if } -3 \le z < -2,\\
&= TSeries[e^{-z^2/2},\ z = -3/2, d = 3] && \text{if } -2 \le z < -1,\\
&= TSeries[e^{-z^2/2},\ z = -1/2, d = 3] && \text{if } -1 \le z < 0,\\
&= TSeries[e^{-z^2/2},\ z = 1/2,\ \ d = 3] && \text{if } 0 \le z < 1,\\
&= TSeries[e^{-z^2/2},\ z = 3/2,\ \ d = 3] && \text{if } 1 \le z < 2,\\
&= TSeries[e^{-z^2/2},\ z = 5/2,\ \ d = 3] && \text{if } 2 \le z < 3,\\
&= TSeries[e^{-z^2/2},\ z = 7/2,\ \ d = 4] && \text{if } 3 \le z < 4,\\
&= 0, && \text{otherwise.}
\end{aligned} \tag{2.5}
$$

Notice that the degree of the first and the eighth piece is 4 to avoid $\varphi_{8p}(z) < 0$. $E(\varphi_{8p}) = 0$, and $V(\varphi_{8p}) \approx 0.99985$. Due to the larger interval, the variance is closer to 1 than the variance for

φ_{6p}. If we truncate the PDF of the standard normal to the region (–4, 4) and renormalize it, then its variance is ≈ 0.99893.

To find a MOP approximation of the PDF of the $N(\mu, \sigma^2)$ distribution, where μ and $\sigma > 0$ are constants, we exploit the fact that MOP functions are invariant under linear transformations. Thus, if $f(x)$ is a MOP function, then $f(ax + b)$ is also a MOP function. If $Z \sim N(0, 1)$, its PDF is approximated by a MOP function $\varphi_p(x)$, and $X = \sigma Z + \mu$, then $X \sim N(\mu, \sigma^2)$, and a MOP approximation of the PDF of X is given by $\xi(x) = (1/\sigma)\varphi_p((x - \mu)/\sigma)$.

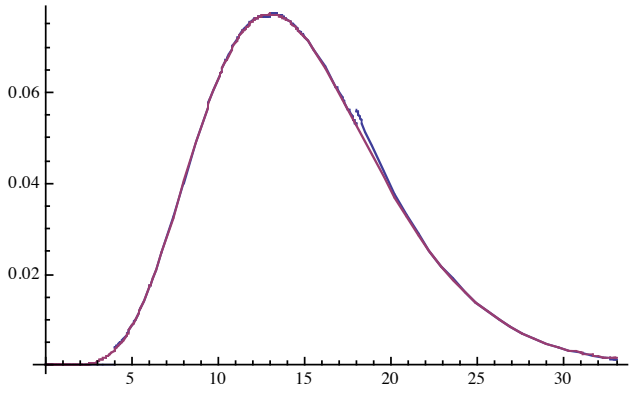
Example 2.2. Consider the chi-square distribution with 15 degrees of freedom, denoted by $\chi^2(15)$. Let $\chi_{15}(y)$ denote the PDF of this distribution. The 0.50 percentile of χ_{15} is ≈ 4.6, and 99.5 percentile is 32.8. Therefore, we will find a MOP approximation on the interval (4, 33). The mode of $\chi_{15}(y)$ is 13, the inflection points are at $13 - \sqrt{26}$ and $13 + \sqrt{26}$ Therefore, we split the domain of $\chi_{15}(y)$ into 4 pieces, (4, 8], (8, 13], (13, 18], and (18, 33), and then we find a MOP approximation of each piece. A 4-piece, 3-degree, MOP approximation of χ_{15} on the interval (4, 33), denoted by χ_{4p}, is as follows:

$$\begin{aligned}
\chi_{4p}(y) &= TSeries[\chi_{15}(y), y = 6, \quad d = 3] && \text{if } 4 < y < 8,\\
&= TSeries[\chi_{15}(y), y = 21/2, \quad d = 3] && \text{if } 8 \le y < 13,\\
&= TSeries[\chi_{15}(y), y = 31/2, \quad d = 3] && \text{if } 13 \le y < 18,\\
&= TSeries[\chi_{15}(y), y = 51/2, \quad d = 3] && \text{if } 18 \le y < 33, \text{ and}\\
&= 0, && \text{otherwise.}
\end{aligned} \tag{2.6}$$

The degree of the approximation in each piece was determined by graphing χ_{15} and χ_{4p}, making sure $\chi_{4p} \ge 0$, and ensuring that we have a good approximation.

A graph of the MOP approximation χ_{4p} overlaid on the actual PDF χ_{15} is shown in Figure 3. The mean of χ_{4p} is ≈ 14.94 (compared to 15 for χ_{15}) and its variance ≈ 27.71 (compared to 30 for χ_{15}). The errors in the mean and variance are mostly the result of truncation of the right and left tails of the distribution. If we restrict the χ_{15} PDF to the interval (4, 33), normalize it so that it is a PDF, then its mean is ≈ 14.93, and its variance is ≈ 27.77.

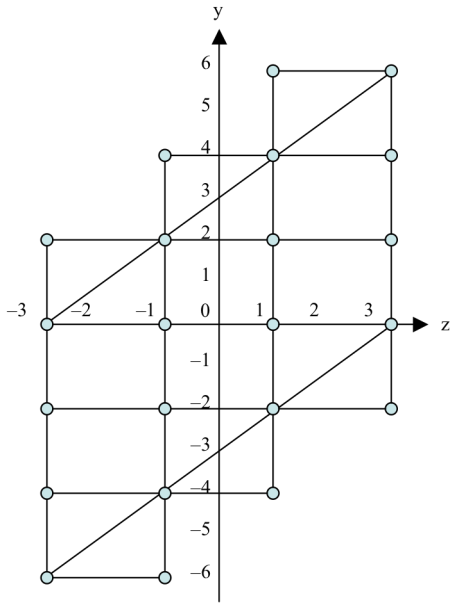
Figure 3. The PDF of $\chi^2(15)$ (in red) and its polynomial approximation χ_{4p} (in blue).



Example 2.3. Consider the CLG distribution $Y|z \sim N(z, 1)$, where $Z \sim N(0, 1)$. Let $\varphi(z, y) = (1/\sqrt{2\pi})e^{-(y-z)^2/2}$ denote the PDF of $N(z, 1)$. Using three standard deviations, the conditional PDF of Y is on the two-dimensional space $-3 \le z \le 3, z - 3 \le y \le z + 3$. This space is covered by twelve squares as shown in Figure 4. We find a MOP approximation for each square at the mid-point in the square. Let $\psi(z, y)$ denote a 12-piece, 14-degree MOP approximation of $\varphi(z, y)$. Then,

$$
\begin{aligned}
\psi(z, y) &= TSeries[e^{-(y-z)^2/2}, z = -2, d_z = 7, y = -5, d_y = 7] && \text{if } -3 \le z \le -1,\ -6 \le y < -4;\\
&= TSeries[e^{-(y-z)^2/2}, z = -2, d_z = 7, y = -3, d_y = 7] && \text{if } -3 \le z \le -1,\ -4 \le y < -2;\\
&= TSeries[e^{-(y-z)^2/2}, z = -2, d_z = 7, y = -1, d_y = 7] && \text{if } -3 \le z \le -1,\ -2 \le y < -0;\\
&= TSeries[e^{-(y-z)^2/2}, z = -2, d_z = 7, y = 1, d_y = 7] && \text{if } -3 \le z \le -1,\ 0 \le y \le 2;\\
&= TSeries[e^{-(y-z)^2/2}, z = 0, d_z = 7, y = -3, d_y = 7] && \text{if } -1 < z \le 1,\ -4 \le y < -2;\\
&= TSeries[e^{-(y-z)^2/2}, z = 0, d_z = 7, y = -1, d_y = 7] && \text{if } -1 < z \le 1,\ -2 \le y < 0;\\
&= TSeries[e^{-(y-z)^2/2}, z = 0, d_z = 7, y = 1, d_y = 7] && \text{if } -1 < z \le 1,\ 0 \le y < 2;\\
&= TSeries[e^{-(y-z)^2/2}, z = 0, d_z = 7, y = 3, d_y = 7] && \text{if } -1 \le z \le 1,\ 2 \le y \le 4;\\
&= TSeries[e^{-(y-z)^2/2}, z = 2, d_z = 7, y = -1, d_y = 7] && \text{if } 1 < z \le 3,\ -2 \le y < 0;\\
&= TSeries[e^{-(y-z)^2/2}, z = 2, d_z = 7, y = 1, d_y = 7] && \text{if } 1 < z \le 3,\ 0 \le y < 2;\\
&= TSeries[e^{-(y-z)^2/2}, z = 2, d_z = 7, y = 3, d_y = 7] && \text{if } 1 < z \le 3,\ 2 \le y < 4;\\
&= TSeries[e^{-(y-z)^2/2}, z = 2, d_z = 7, y = 5, d_y = 7] && \text{if } 1 \le z \le 3,\ 4 \le y \le 6;\\
&= 0, && \text{otherwise} \qquad (2.7)
\end{aligned}
$$

Figure 4. The region over which a MOP approximation of $\varphi(z, y)$ is computed.



In (2.7), we are using the two-dimensional Taylor series approximation of $\varphi(z, y)$. Notice that for any function $f(x, y)$, the two dimensional Taylor series expansion of f at the point (a, b), $TSeries[f(x, y), x = a, d_x, y = b, d_y]$ can be written as a product of two one-dimensional Taylor series expansions as follows:

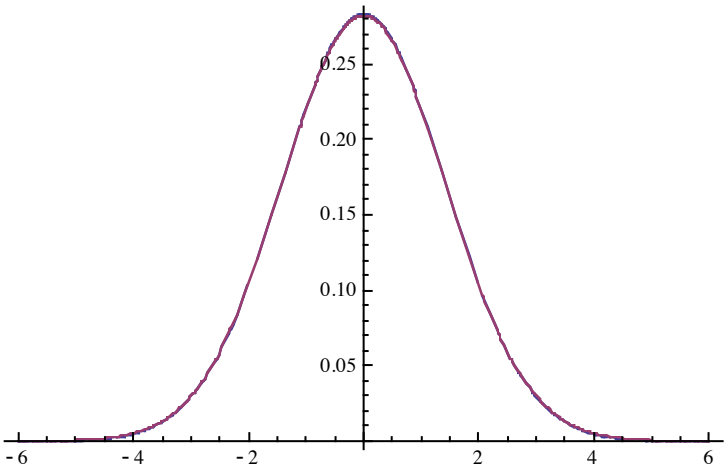
$$TSeries[f(x, b), x = a, d_x]\ TSeries[f(a, y), y = b, d_y].$$

Thus, the two-dimensional Taylor series expansion in (2.7) yields a two-dimensional MOP function by definition in (2.2).

Let $\psi_u(z,y)$ denote the un-normalized MOP approximation of the PDF of $Y|z$. Since $h(z) = \int \psi_u(z,y)\, dy$ is a function of z, we cannot normalize the MOP approximation in the usual way since $(1/h(z))\, \psi_u(z, y)$ may not be a MOP function. However, we can partially normalize it as follows. Let $\varphi_{6p}(z)$ described in (2.4) denote the (normalized) 6-piece, 3-degree MOP approximation of the PDF of $N(0, 1)$ on $(-3, 3)$. Then $\varphi_{6p}(z)\ \psi_u(z, y)$ represents the un-normalized joint PDF of (Z, Y). We compute the normalization constant $c = \iint \varphi_{6p}(z)\ \psi_u(z, y)\ dy\ dz$, and $\psi(z, y) = c^{-1}\ \psi_u(z, y)$ represents the partially normalized PDF of $Y|z$ in (2.7).

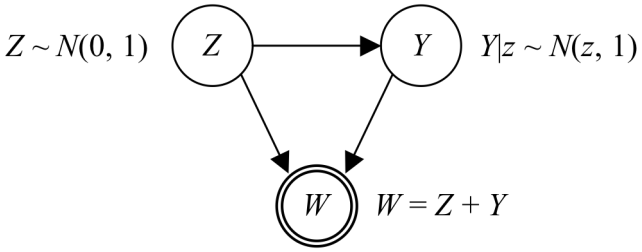
As a first check on the quality of the MOP approximation of $Y|z$, we computed the MOP approximation of the marginal distribution of Y as follows: $\psi'(y) = \int \varphi_{6p}(z)\ \psi(z, y)\ dz$. $\psi'(y)$ is computed as an 8-piece, 7-degree MOP function. The exact marginal of Y is $N(0, 2)$. A plot of $\psi'(y)$ overlaid on the PDF of $N(0, 2)$ is shown in Figure 5. Also, $\int \psi'(y)\, dy = 1$, the mean of $\psi' = 0$, and the variance of $\psi' \approx 1.970$.

Figure 5. The graph of ψ' (in blue) overlaid on the PDF of $N(0, 2)$ (in red).



As a second check on the quality of the MOP approximation of the PDF of $N(z, 1)$, consider the Bayesian network as shown in Figure 6 that includes W with a deterministic conditional $W = Z + Y$.

Figure 6. A Bayesian network with a sum deterministic function

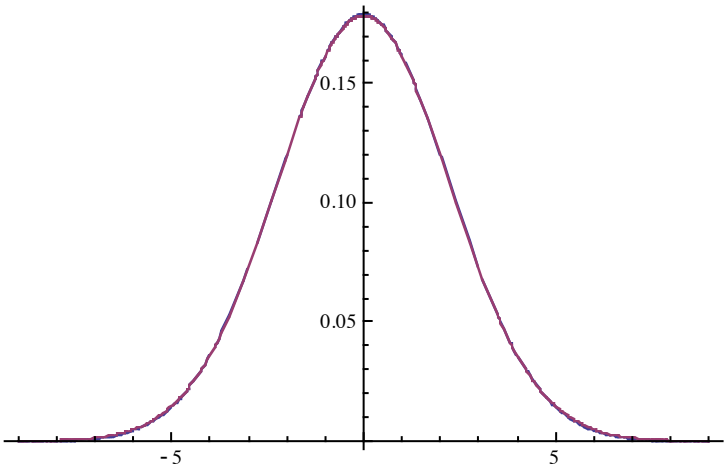


Suppose we use $\varphi_{6p}(z)$ as a MOP approximation for the PDF of Z, and $\psi(z, y)$ as a MOP approximation of the conditional PDF of Y. The marginal distribution of W is given by the convolution formula:

$$\omega(w) = \int \varphi_{6p}(z)\, \psi(z, w - z)\, dz \qquad (2.8)$$

Notice that $\omega(w)$ is a MOP function. The integration in (2.8) was done in Mathematica®. ω is computed as a 33-piece, 18-degree MOP function. The exact joint distribution of (Z, Y) is bivariate normal with parameters $\mu_Z = \mu_Y = 0$, $\sigma_Z^2 = 1$, $\sigma_Y^2 = 2$, $\sigma_{ZY} = 1$. Therefore, the exact marginal distribution of W is $N(0, 5)$. A graph of ω overlaid on the graph of the PDF of $N(0, 5)$ is shown in Figure 7. Also, $\int \omega(w)\, dw = 1$, $E(\omega) = 0$, and $V(\omega) \approx 4.892$.

Figure 7. The graph of ω (in blue) overlaid on the graph of the PDF of $N(0, 5)$



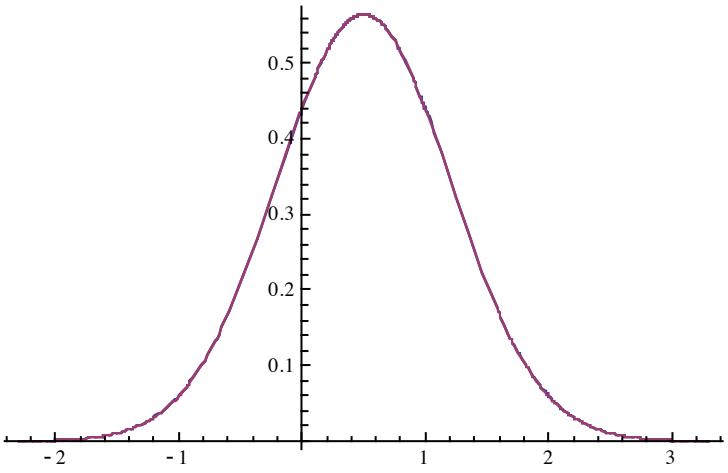
As a third check on the quality of the MOP approx. of the PDF of $N(z, 1)$, suppose we observe $Y = 1$, and we wish to compute the posterior marginal of Z. Suppose $\varphi_{6p}(z)$ denotes the MOP approximation of the prior PDF of Z, and $\psi(z, y)$ the conditional PDF of Y given z. The un-normalized posterior marginal of Z is given by $\varphi_{6p}(z)\ \psi(z, 1)$, and the normalized posterior marginal of Z is given by

$$\zeta(z) = c^{-1}\ \varphi_{6p}(z)\ \psi(z, 1), \text{ where } c = \int \varphi_{6p}(z)\ \psi(z, 1)\ dz. \tag{2.9}$$

The exact posterior marginal of Z is $N(½, ½)$. A plot of the function $\zeta(z)$ overlaid on the exact PDF of $N(½, ½)$ is shown in Figure 8. $E(\zeta) \approx 0.4999$, and $V(\zeta) \approx 0.4992$.

Based on the results of the three checks described above, we conclude that the partial normalization of the MOP approximation of Y given z does not introduce any problems. In the first two checks, the results are exactly normalized as a result of partial normalization. In the third check, since we have to normalize the posterior marginal in light of the observation, the partial normalization is sufficient.

Figure 8. A plot of $\zeta(z)$ (in blue) overlaid on the exact PDF of $N(½, ½)$ (in red)



As described for the normal distribution, one can extend the MOP approximation of $N(z, 1)$ to an arbitrary CLG distribution $W|x \sim N(ax + b, d^2)$ in two dimensions. Suppose, for example, we wish to have the MOP approximation of PDF of $W|x \sim N(ax + b, d^2)$, where $X \sim N(\mu, \sigma^2)$ and μ, $\sigma > 0$, a, b, and $d > 0$ are some specified constants . Let $\psi(z, y)$ denote the MOP approximation of PDF of $Y|z \sim N(z, 1)$, where $Z \sim N(0, 1)$. Suppose $X = \sigma Z + \mu$, and $W|x = d(Y|z) + (ax + b - (dx/\sigma)) + d\mu/\sigma)$. First, notice that both are linear transformations. Second, since $Z \sim N(0, 1)$, $X \sim N(\mu, \sigma^2)$. Third, $Y|z = Y|(x - \mu)/\sigma \sim N((x - \mu)/\sigma, 1)$. Therefore, $W|x = d(Y|z) + (ax + b - dx/\sigma + d\mu/\sigma) \sim N(ax + b, d^2)$. Finally, a MOP approximation of the PDF of $W|x$ is given by:

$$\omega(x, w) = \frac{1}{d}\psi\left(\frac{x-\mu}{\sigma}, \frac{w-(ax+b-dx/\sigma+d\mu/\sigma)}{d}\right) \qquad (2.10)$$

3 The Extended Shenoy-Shafer Architecture

In this section, we sketch the extended Shenoy-Shafer architecture [Shenoy and West 2009a] for representing and solving hybrid BNs with deterministic variables. The Shenoy-Shafer architecture [Shenoy and Shafer 1990] was initially proposed for computing marginals in discrete Bayesian networks. It was extended by Moral *et al.* [2001] to include continuous variables for propagation of mixtures of truncated exponentials. Cobb and Shenoy [2005a] extended it further to include linear deterministic variables. Cinicioglu and Shenoy [2009] extended it further to include linear and non-linear deterministic functions to define arc reversals. They propose the use of Dirac delta functions for representing conditionals of deterministic functions. Finally, Shenoy and West [2009a] extended it further to include mixed potentials to

compute marginals in Bayesian networks with differentiable deterministic functions. Since an architecture to propagate the MOP potentials is essential, we will sketch it here. Most of the material in this section is taken from Shenoy and West [2009a].

Variables and States. We are concerned with a finite set V of variables. Each variable $X \in V$ is associated with a set Ω_X of its possible states. If Ω_X is a finite set or countably infinite, we say X is *discrete*, otherwise X is *continuous*. We will assume that the state space of continuous variables is the set of real numbers (or some subset of it), and that the state space of discrete variables is a set of symbols (not necessarily real numbers). If $r \subseteq V$, $r \neq \varnothing$, then $\Omega_r = \times\{\Omega_X \mid X \in r\}$. If $r = \varnothing$, we will adopt the convention that $\Omega_\varnothing = \{\blacklozenge\}$.

In a BN, each variable has a conditional distribution function for each state of its parents. A conditional distribution function associated with a continuous variable is said to be *deterministic* if the variances (for each state of its parents) are all zeros. For simplicity, henceforth, in the context of a BN representation, we will refer to continuous variables with non-deterministic conditionals as *continuous*, and continuous variables with deterministic conditionals as *deterministic*. In a BN, discrete variables are denoted by rectangular-shaped nodes, continuous by oval-shaped nodes, and deterministic variables by oval-shaped nodes with a double border.

Projection of States. If $\boldsymbol{x} \in \Omega_r$, $\boldsymbol{y} \in \Omega_s$, and $r \cap s = \varnothing$, then $(\boldsymbol{x}, \boldsymbol{y}) \in \Omega_{r \cup s}$. Thus, $(\boldsymbol{x}, \blacklozenge) = \boldsymbol{x}$. Suppose $\boldsymbol{x} \in \Omega_r$, and suppose $s \subseteq r$. Then the *projection* of $\boldsymbol{x}$ to s, denoted by $\boldsymbol{x}^{\downarrow s}$, is the state of s obtained from $\boldsymbol{x}$ by dropping states of $r \backslash s$. Thus, $(w, x, y, z)^{\downarrow\{W, X\}} = (w, x)$, where $w \in \Omega_W$, and $x \in \Omega_X$. If $s = r$, then $\boldsymbol{x}^{\downarrow s} = \boldsymbol{x}$. If $s = \varnothing$, then $\boldsymbol{x}^{\downarrow s} = \blacklozenge$.

Discrete Potentials. In a BN, the conditional probability functions associated with the variables are represented by functions called *potentials*. If A is discrete, it is associated with conditional probability mass functions, one for each state of its parents. The conditional probability mass functions are represented by functions called *discrete potentials*.

Suppose $r \subseteq V$ is such that it contains a discrete variable. A *discrete* potential α for r is a function $\alpha: \Omega_r \rightarrow [0, 1]$. The values of discrete potentials are probabilities.

Although the domain of the function α is Ω_r, for simplicity, we will refer to r as the *domain* of α. Thus, the domain of a potential representing the conditional probability function associated with some variable X in a BN is always the set $\{X\} \cup pa(X)$, where $pa(X)$ denotes the set of parents of X in the BN graph.

Notice that a discrete potential can have continuous variables in its domain, but if so, it has to include a discrete variable, and its values are in units of probability. For example, if A is a discrete variable with two states a_1 and a_2, the values of a discrete potential for A, $\alpha(a_1) = ½$ and $\alpha(a_2) = ½$, are in units of probability.

Density Potentials. If Z is continuous, then it is usually associated with a *density* potential. Suppose $r \subseteq V$ is such that it contains a continuous variable. A density potential ζ for r is a function $\zeta: \Omega_r \rightarrow \mathbb{R}^+$, where $\mathbb{R}^+$ is the set of non-negative real numbers. The values of density potentials are probability densities.

Notice that a density potential can have discrete variables in its domain, but if so, it has to include a continuous variable, and its values are in units of density. For example, if Z is a continuous variable with the standard normal distribution, then the values of a continuous potential for Z, $\zeta(z) = (1/\sqrt{2\pi})e^{-z^2/2}$, are in units of density.

Dirac Delta Functions. Deterministic variables have conditional distributions containing equations. We will represent such functions by *Dirac* potentials based on the Dirac delta function δ.

$\delta: \mathbb{R} \rightarrow \mathbb{R}^+$ is called a *Dirac delta function* if $\delta(x) = 0$ if $x \neq 0$, and $\int \delta(x)\, dx = 1$. δ is not a proper function since the value of the function at 0 doesn't exist (i.e., is not finite). It can be regarded as a limit of a certain sequence of functions (such as, e.g., the Gaussian density function with mean 0 and variance σ^2 in the limit as $\sigma \rightarrow 0$). However, it can be used as if it were a proper function for practically all our purposes without getting incorrect results. It was first defined by Dirac [1927].

As defined above, the value $\delta(0)$ is undefined, i.e., ∞, when considered as probability density. We argue that we can interpret the value $\delta(0)$ as probability 1. Consider the normal PDF with mean 0 and variance σ^2. Its moment generating function (MGF) is $M(t) = e^{\sigma^2 t}$. In the limit as $\sigma \rightarrow 0$, $M(t) = 1$. Now, $M(t) = 1$ is the MGF of the distribution $X = 0$ with probability 1. Therefore, we can interpret the value $\delta(0)$ (in units of density) as probability 1 at the location $x = 0$.

Some basic properties of the Dirac delta function are as follows [Dirac 1927, Dirac 1958, Hoskins 1979, Kanwal 1998, Saichev and Woyczynski 1997, Khuri 2004].

1. If $f(x)$ is any function that is continuous in the neighborhood of a, then $\int f(x)\, \delta(x - a)\, dx = f(a)$.
2. $\int \delta(x - h(u, v))\, \delta(y - g(v, w, x))\, dx = \delta(y - g(v, w, h(u, v)))$.
3. If $g(x)$ has real (non-complex) zeros at $a_1, \ldots, a_n$, and is differentiable at these points, and $g'(a_i) \neq 0$ for $i = 1, \ldots, n$, then $\delta(g(x)) = \Sigma_i\, \delta(x - a_i)/|g'(a_i)|$.
4. Suppose continuous variable X has PDF $f_X(x)$ and $Y = g(X)$. Then Y has PDF $f_Y(y) = \int f_X(x)\, \delta(y - g(x))\, dx$.

A more extensive list of properties of the Dirac delta function that is relevant for uncertain reasoning can be found in [Cinicioglu and Shenoy 2009].

Dirac Potentials. Suppose $t = r \cup s$ is a set of variables containing some discrete variables r and some continuous variables s. We assume $s \neq \emptyset$. A *Dirac* potential ξ for t is a function $\xi: \Omega_t \rightarrow \mathbb{R}^+$ such that $\xi(\boldsymbol{r}, \boldsymbol{s})$ is of the form:

$$\xi(\boldsymbol{r}, \boldsymbol{s}) = \Sigma\{p_{\boldsymbol{r},i}\,\delta(z - g_{\boldsymbol{r},i}(\boldsymbol{s}^{\downarrow(s\setminus\{Z\})}))\,|\,\boldsymbol{r} \in \Omega_r, \text{ and } i = 1, \ldots, n_{\boldsymbol{r}}\}, \tag{3.1}$$

where $\boldsymbol{r} \in \Omega_r$, $\boldsymbol{s} \in \Omega_s$, $Z \in s$ is a continuous or deterministic variable, $z \in \Omega_Z$, $\delta(z - g_{\boldsymbol{r},i}(\boldsymbol{s}^{\downarrow(s\setminus\{Z\})}))$ are Dirac delta functions, $p_{\boldsymbol{r},i}$ are probabilities for all $i = 1, \ldots, n_{\boldsymbol{r}}$, and $n_{\boldsymbol{r}}$ is a positive integer. Here, we are assuming that continuous or deterministic variable Z is a weighted sum of deterministic functions $g_{\boldsymbol{r},i}(\boldsymbol{s}^{\downarrow(s\setminus\{Z\})})$ of the other continuous variables in s, with probability weights $p_{\boldsymbol{r},i}$, and that the nature of the deterministic functions and weights may depend on the state $\boldsymbol{r}$ of the discrete variables in r, or on some latent index i.

Suppose Y is a deterministic variable with continuous parent X, and suppose that the deterministic relationship is $Y = X^2$. This conditional distribution is represented by the Dirac potential $\delta(y - x^2)$ for $\{X, Y\}$. Here $r = \emptyset$, and $n_\blacklozenge = 1$, in the definition in (3.1).

A second example of a Dirac potential for $\{A, X\}$ where A is discrete with states a and na, and X is continuous is as follows. $\xi(a, x) = (½)\delta(x - 1) + (½)\delta(x - 2)$, $\xi(na, x) = (1/3)\delta(x - 1) + (1/3)\delta(x - 2) + (1/3)\delta(x - 3)$. Here, $r = \{A\}$, $n_a = 2$, $n_{na} = 3$, in the definition in (3.1). ξ can be interpreted as follows: ξ is the conditional for X given A; if $A = a$, then $X = 1$ with probability ½, and $X = 2$ with probability ½; if $A = na$, then $X = 1$ with probability 1/3, $X = 2$ with probability 1/3, and $X = 3$ with probability 1/3.

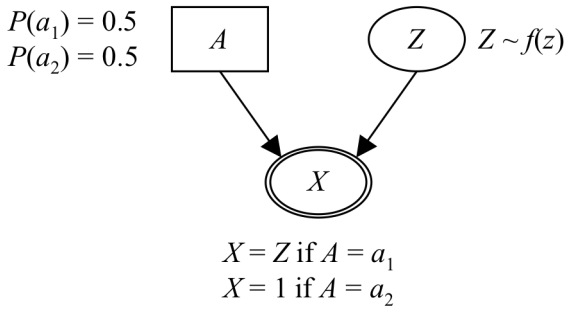
A third example of a Dirac potential for $\{Z, X\}$, where Z and X are both continuous variables is as follows: $\xi(z, x) = (1/2)\delta(x - z) + (1/2)\delta(x - 1)$. Here, $r = \emptyset$, $n_\blacklozenge = 2$, $p_1 = ½$ and $p_2 = ½$ in the definition in (3.1). The two probability weights correspond to some latent index (the states of a marginalized discrete variable). This Dirac potential can be interpreted as a conditional for X given $\{Z\}$ as follows: Given $Z = z$, $X = z$ with probability ½, and $X = 1$ with probability ½.

Continuous Potentials. Both density and Dirac potentials are special instances of a broader class of potentials called continuous potentials. Suppose $t \subseteq V$ is such that it contains a continuous variable. Then, a *continuous potential* ξ for t is a function $\xi: \Omega_t \rightarrow \mathbb{R}^+$. The values of continuous potentials are in units of density. For example, consider a continuous variable X with a mixed distribution: a probability of 0.5 at $X = 1$, and a probability density of 0.5 $f(x)$, where $f(x)$ is a PDF. This mixed distribution can be represented by a continuous potential ξ for $\{X\}$ as follows: $\xi(x) = 0.5\,\delta(x - 1) + 0.5 f(x)$. Notice that $\int \xi(x)\,dx = 0.5 \int \delta(x-1)\,dx + 0.5 \int f(x)\,dx = 0.5 + 0.5 = 1$.

Consider the BN in Figure 9. A is discrete (with two states, a_1 and a_2), Z is continuous, and X is deterministic. Let α denote the discrete potential for {A}. Then $\alpha(a_1) = 0.5$, $\alpha(a_2) = 0.5$. Let ζ

denote the density potential for $\{Z\}$. Then $\zeta(z) = f(z)$. Let ξ denote the Dirac potential for $\{A, Z, X\}$. Then $\xi(a_1, z, x) = \delta(x - z)$, and $\xi(a_2, z, x) = \delta(x - 1)$.

Figure 9. A hybrid BN with a discrete, a continuous and a deterministic variable.



Mixed Potentials. To reason with hybrid models, we need to define mixed potentials. A mixed potential has two parts, the first part is a discrete potential and the second part is a continuous potential. Formally, suppose α is a discrete potential for r. Then a mixed potential representation of α is $\mu_1 = (\alpha, \iota_c)$ for r, where ι_c denotes the *identity* continuous potential for the empty set, $\iota_c(\blacklozenge) = 1$. Suppose ζ is a continuous potential for s. Then, a mixed potential representation of ζ is $\mu_2 = (\iota_d, \zeta)$ for s, where ι_d denotes the *identity* discrete potential for the empty set, $\iota_d(\blacklozenge) = 1$. The difference between ι_c and ι_d is that the former is in unit of density, whereas the latter is in unit of probability. Mixed potentials can have non-vacuous discrete and continuous parts. Thus $\mu_3 = (\alpha, \zeta)$ is a mixed potential for $r \cup s$. Such a mixed potential would be the result of combining μ_1 and μ_2, which we will define next. The main idea behind mixed potentials is to represent the nature (discrete or continuous) of potentials.

Combination of Potentials. Suppose α is a discrete or continuous potential for some subset a of variables, and β is a discrete or continuous potential for some subset b. Then the *combination* of α and β, denoted by $\alpha \otimes \beta$, is the potential for $a \cup b$ obtained from α and β by pointwise multiplication, i.e.,

$$(\alpha \otimes \beta)(\boldsymbol{x}) = \alpha(\boldsymbol{x}^{\downarrow a})\, \beta(\boldsymbol{x}^{\downarrow b}), \text{ for all } \boldsymbol{x} \in \Omega_{a \cup b}. \tag{3.2}$$

If α and β are both discrete potentials, then $\alpha \otimes \beta$ is a discrete potential, and if α and β are both continuous potentials, then $\alpha \otimes \beta$ is a continuous potential. The definition of combination in (3.2) is valid also if α is discrete and β is continuous and vice-versa, and will be used when we define marginalization of mixed potentials. However, the nature of the potential $\alpha \otimes \beta$ when α is discrete and β is continuous (or vice-versa) will not arise in the combination operation. We will use mixed potentials, and as we will see, a combination of mixed potentials avoids such

combinations. However, the combination defined in (3.2) will arise in the marginalization operation for mixed potentials, and we will describe the nature of the combination at that point.

Combination of Mixed Potentials. Suppose $\mu_1 = (\alpha_1, \zeta_1)$, and $\mu_2 = (\alpha_2, \zeta_2)$ are two mixed potentials with discrete parts α_1 for r_1 and α_2 for r_2, respectively, and continuous parts ζ_1 for s_1 and ζ_2 for s_2, respectively. Then, the combination $\mu_1 \otimes \mu_2$ is a mixed potential for $r_1 \cup s_1 \cup r_2 \cup s_2$ given by

$$\mu_1 \otimes \mu_2 = (\alpha_1 \otimes \alpha_2, \zeta_1 \otimes \zeta_2). \tag{3.3}$$

Since $\alpha_1 \otimes \alpha_2$ is a discrete potential and $\zeta_1 \otimes \zeta_2$ is a continuous potential, the definition of combination of mixed potentials in (3.3) is consistent with the definition of mixed potentials. If $\mu_1 = (\alpha, \iota_c)$ represents the discrete potential α for r, and $\mu_2 = (\iota_d, \zeta)$ represents the continuous potential for s, then $\mu_1 \otimes \mu_2 = (\alpha, \zeta)$ is a mixed potential for $r \cup s$.

Since combination is pointwise multiplication, and multiplication is commutative, combination of potentials (discrete or continuous) is commutative ($\alpha \otimes \beta = \beta \otimes \alpha$) and associative ($(\alpha \otimes \beta) \otimes \gamma = \alpha \otimes (\beta \otimes \gamma)$). Since the combination of mixed potentials is defined in terms of combination of discrete and continuous potentials, each of which is commutative and associative, combination of mixed potentials is also commutative and associative.

Marginalization of Potentials. The definition of marginalization depends on whether the variable being marginalized is discrete or continuous. We marginalize discrete variables by addition and continuous variables by integration. Integration of potentials containing Dirac delta functions is done using properties of Dirac delta functions. In addition, after marginalization, the nature of a potential could change, e.g., from continuous to discrete (if the domain of the marginalized potential contains only discrete variables) and from discrete to continuous (if the domain of the marginalized potential contains only continuous variables). We will make this more precise when we define marginalization of mixed potentials.

Suppose α is a discrete or continuous potential for a, and suppose X is a discrete variable in a. Then the *marginal* of α by deleting X, denoted by α^{-X}, is the potential for $a \backslash \{X\}$ obtained from α by addition over the states of X, i.e.,

$$\alpha^{-X}(\boldsymbol{y}) = \Sigma\{\alpha(x, \boldsymbol{y}) \mid x \in \Omega_X\} \text{ for all } \boldsymbol{y} \in \Omega_{a \backslash \{X\}}. \tag{3.4}$$

If X is a continuous variable in a, then the marginal of α by deleting X is obtained by integration over the state space of X, i.e.,

$$\alpha^{-X}(\boldsymbol{y}) = \int \alpha(x, \boldsymbol{y})\, dx \text{ for all } \boldsymbol{y} \in \Omega_{a \backslash \{X\}}. \tag{3.5}$$

If ξ is a discrete or continuous potential for $\{X\} \cup pa(X)$ representing the conditional distribution for X in a BN, then ξ^{-X} is an identity potential for $pa(A)$, i.e., a potential whose

values are all ones. The nature of this potential (discrete or continuous) will be described when we discuss marginalization of mixed potentials.

If we marginalize a discrete or continuous potential by deleting two (or more) variables from its domain, then the order in which the variables are deleted does not matter, i.e., $(\alpha^{-A})^{-B} = (\alpha^{-B})^{-A} = \alpha^{-\{A, B\}}$.

If α is a discrete or continuous potential for a, β is a discrete or continuous potential for b, $A \in a$, and $A \notin b$, then $(\alpha \otimes \beta)^{-A} = \alpha^{-A} \otimes \beta$. This is a key property of combination and marginalization that allows local computation [Shenoy and Shafer 1990]. We will refer to this property as *local computation*.

Marginalization of Mixed Potentials. Mixed potentials allow us to represent the nature of potentials, and marginalization of mixed potentials allows us to represent the nature of the marginal. Suppose $\mu = (\alpha, \zeta)$ is a mixed potential for $r \cup s$ with discrete part α for r, and continuous part ζ for s. Let C denote the set of continuous variables, and let D denote the set of discrete variables. The marginal of μ by deleting $X \in r \cup s$, denoted by μ^{-X}, is defined as follows.

$$\mu^{-X} = \begin{cases} (\alpha^{-X}, \zeta) & \text{if } X \in r, X \notin s, \text{ and } r \backslash \{X\} \not\subset C; \quad (3.6) \\ (\iota_d, \alpha^{-X} \otimes \zeta) & \text{if } X \in r, X \notin s, \text{ and } r \backslash \{X\} \subseteq C; \quad (3.7) \\ (\alpha, \zeta^{-X}) & \text{if } X \notin r, X \in s, \text{ and } s \backslash \{X\} \not\subset D; \quad (3.8) \\ (\alpha \otimes \zeta^{-X}, \iota_c) & \text{if } X \notin r, X \in s, \text{ and } s \backslash \{X\} \subseteq D; \quad (3.9) \\ ((\alpha \otimes \zeta)^{-X}, \iota_c) & \text{if } X \in r, X \in s, \text{ and } (r \cup s) \backslash \{X\} \subseteq D; \quad (3.10) \\ (\iota_d, (\alpha \otimes \zeta)^{-X}) & \text{if } X \in r, X \in s, \text{ and } (r \cup s) \backslash \{X\} \not\subset D. \quad (3.11) \end{cases}$$

Some comments about the definition of marginalization of mixed potentials are as follows. First, if the variable being deleted belongs only to one part (discrete or continuous, as in cases 3.6–3.9), then the local computation property allow us to delete the variable from that part only leaving the other part unchanged. If the variable being deleted belongs to both parts (as in cases 3.10–3.11), then we first need to combine the two parts before deleting the variable. Second, when we have only continuous variables left in a discrete potential after marginalization, we move the potential to the continuous part (3.7) and when we only have discrete variables left, we move the potential to the discrete part (3.9), otherwise we don't change the nature of the marginalized potentials (3.6 and 3.8). In cases 3.10 and 3.11, when we have to combine the discrete and continuous potentials before marginalizing X, if only discrete variables are left, then we have to classify it as a discrete potential (3.10), and if we have only continuous variables left, then we have to classify it as a continuous potential (3.11). However, if it has discrete and continuous variables, then it is classified as continuous since the product of probability and density is in units of density.

This completes the sketch of the extended Shenoy-Shafer architecture. Shenoy and West [2009a] solve exactly an extended version of the crop problem [Murphy 1999, Lerner 2002] that has been extended further to include deterministic variables with non-linear functions.

4 Some Examples

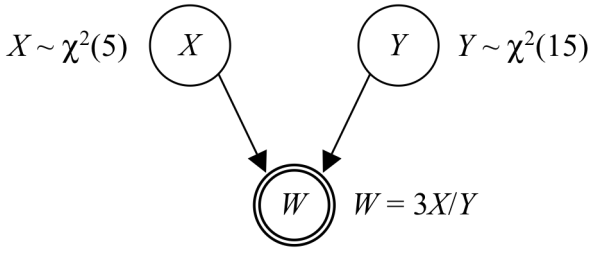
In this section, we describe three small examples to demonstrate the feasibility of using MOP approximations in hybrid BN with deterministic variables. The examples are chosen so that the exact answers are known. This allows us to check how close the MOP approximations are to the exact answers.

In some of these examples, the use of MTE approximations is not possible. Cobb and Shenoy [2005b] propose approximating non-linear deterministic functions by piecewise linear deterministic functions, and then using MTEs. However, they do not report the extent of the errors introduced by using piecewise linear functions as approximations of non-linear deterministic functions.

4.1 Example 1: A Quotient Deterministic Function

Consider a BN as shown in Figure 10. X and Y are independent with $\chi^2(5)$ and $\chi^2(15)$ distributions, respectively, and $W = 3X/Y$. We know from probability theory that W has the F-distribution with 5 and 15 numerator and denominator degrees of freedom, respectively, which we denote by $F(5, 15)$.

Figure 10. A BN with a quotient deterministic function.



In Section 2, we have described a 4-piece, 3-degree MOP approximation of the PDF of $\chi^2(15)$ on the interval (4, 33), which is denoted by $\psi(y)$. We find a MOP approximation of $\chi^2(5)$ in a similar manner. The mode of $\chi^2(5)$ is 3 and $\int_{0.5}^{18} \chi_5(x)\,dx \approx 0.99$. Let $\xi(x)$ denote a 3-piece, 4-degree MOP approximation of $\chi^2(5)$ on the interval (0.5, 18) as follows.

$$\begin{aligned}
\xi(x) &= TSeries[\chi_5(x), x = 7/4, \quad d = 3] && \text{if } 1/2 < x \le 3, \\
&= TSeries[\chi_5(x), x = 17/4, d = 3] && \text{if } 3 < x \le 11/2, \\
&= TSeries[\chi_5(x), x = 47/4, d = 4] && \text{if } 11/2 < x < 18, \text{ and} \\
&= 0 && \text{otherwise,}
\end{aligned} \tag{4.1}$$

where $\chi_5(x)$ denotes the PDF of $\chi^2(5)$. The mean of $\xi(x)$ is ≈ 5.01 (compared to 5 for $\chi_5(x)$) and the variance is ≈ 9.19 (compared to 10 for $\chi_5(x)$).

Let $\omega(x, y, w) = \delta(w - 3x/y)$ denote the Dirac potential associated with W. To find the marginal for W, first we delete X, and then Y. To delete X, we first combine ξ and ω, and then marginalize X from the combination.

$$(\xi \otimes \omega)^{-X}(y, w) = \int \xi(x)\, \delta(w - 3x/y)\, dx = (|y|/3) \int \xi(x)\, \delta(x - wy/3)\, dx = (|y|/3)\, \xi(wy/3)$$

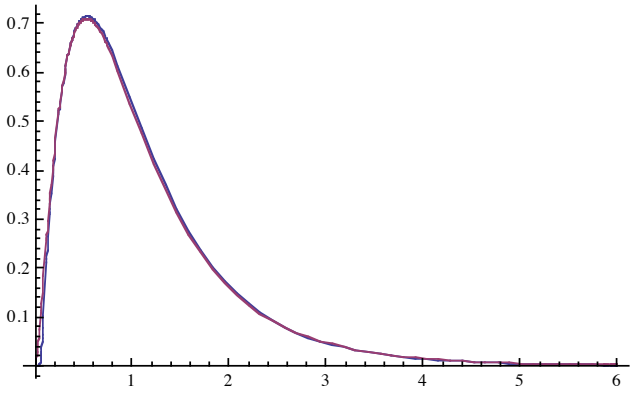
Next, we delete Y by combining $(\xi \otimes \omega)^{-X}$ and ψ, and then marginalizing Y. The result, denoted by ω' is the marginal PDF of W.

$$\omega'(w) = ((\xi \otimes \omega)^{-X} \otimes \psi)^{-Y}(w) = \int_0^\infty (y/3)\, \xi(wy/3)\, \psi(y)\, dy \tag{4.2}$$

Notice that since $\psi(y)$, $\xi(wy/3)$, and $y/3$ are MOP functions that are closed under integration, $\omega'(w)$ is a MOP function.

ω' was computed as a 29-piece, 9-degree polynomial in Mathematica®. A graph of ω' overlaid on the PDF of the $F(5, 15)$ distribution is shown in Figure 11. Notice how well the MOP approximation tracks the exact PDF. The mean of ω' is ≈ 1.150 (compared to 15/13 ≈ 1.154 for $F(5, 15)$) and the variance of ω' is ≈ 0.779 (compared to 1,620/1,859 ≈ 0.871 for $F(5, 15)$). Some of the error in the mean and variance is due to the truncation of the right and left tails of W. The smallest value of W in the MOP approximation is 3(0.5)/33 = 0.045, and the largest is 3(18)/4 = 13.5. If we restrict the $F(5,15)$ distribution to (0.045, 13.5), and normalize it so that it is a PDF, then its mean is 1.155, and its variance is 0.861.

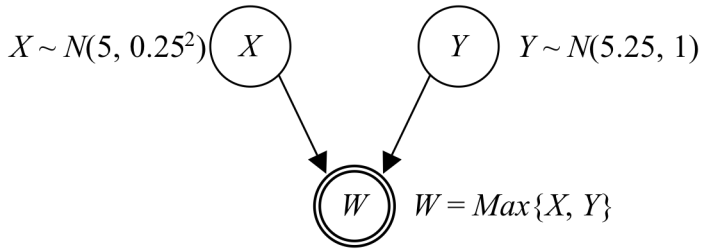
Figure 11. The PDF ω' (in red) overlaid on the PDF of $F(5, 15)$ (in blue).



4.2 Example 2: The Max Deterministic Function

Consider a BN as shown in Figure 12. X and Y are continuous variables and W is deterministic with a non-differentiable function of X and Y, $W = Max\{X, Y\}$.

Figure 12. A BN with a max deterministic function



The conditional associated with W is represented by the Dirac potential $\omega(x, y, w) = \delta(w - Max\{x, y\})$. To compute the marginal PDF of W, we need to evaluate the integral

$$f_W(w) = \int f_X(x) \left(\int f_Y(y)\, \delta(w - Max\{x, y\})\, dy \right) dx \tag{4.3}$$

where $f_W(w)$, $f_X(x)$, and $f_Y(y)$ are the marginal PDF of W, X, and Y, respectively. Since the deterministic function is not differentiable, the integrals in (4.3) cannot be evaluated as written.

One solution to finding the marginal PDF of W is to use theory of order statistics. Let $F_W(w)$, $F_X(x)$, and $F_Y(y)$ denote the marginal cumulative distribution functions (CDF) of W, X, and Y, respectively. Then:

$$F_W(w) = P(W \le w) = P(X \le w,\, Y \le w) = F_X(w)\, F_Y(w). \tag{4.4}$$

Differentiating both sides of the equation with respect to w, we have

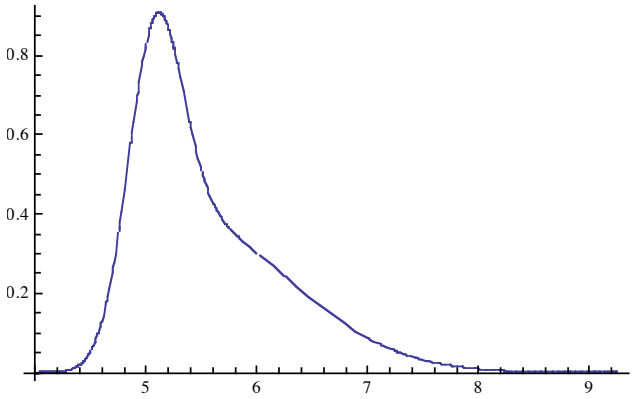
$$f_W(w) = f_X(w)\, F_Y(w) + F_X(w)\, f_Y(w) \tag{4.5}$$

In our example, X and Y have normal PDF, which does not have a closed form CDF. However, using MOP approximations of the normal PDF, we can easily compute a closed form expression for the CDF, which will remain MOP functions. Then, using (4.5), we will have a closed-form MOP approximation for the PDF of W. Assuming we start with the 8-piece, 4-degree MOP approximation φ_{8p} of $N(0, 1)$ on the interval $(-4, 4)$ as described in (2.5), we can find a MOP approximation of the PDF of $N(5, 0.25^2)$ and $N(5.25, 1)$ as discussed in Section 2 as follows.

$$\xi(x) = 4\, \varphi_{8p}(4(x - 5)),$$
$$\psi(y) = \varphi_{8p}(y - 5.25)$$

Next we find the MOP approximations of the CDF of X and Y, and then the MOP approximation of the PDF of W using (4.5). A graph of the MOP approximation of $f_W(w)$ is shown in Figure 13.

Figure 13. A graph of the MOP approximation of the PDF of W.



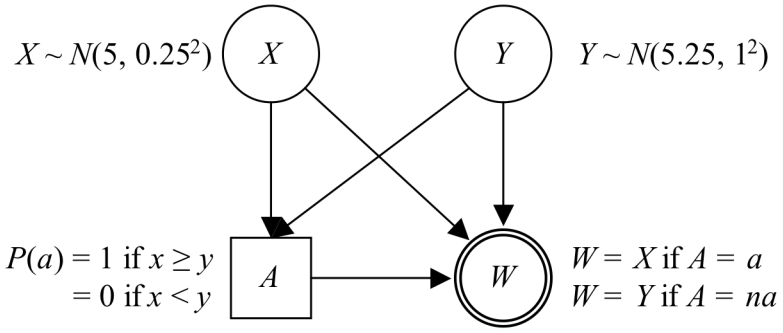
The mean and variance of the MOP approximation of f_W are computed as 5.5484 and 0.4574. Clark [1961] provides formulae for exact computation of the mean and variance of the max of two normals as follows:

$E(W) = E(X)\, F_Z(b) + E(Y)\, F_Z(-b) + a\, f_Z(b)$,
$E(W^2) = (E(X)^2 + V(X))\, F_Z(b) + (E(Y)^2 + V(Y))\, F_Z(-b) + (E(X) + E(Y))\, a\, f_Z(b)$, where
$a^2 = V(X) + V(Y) - 2C(X, Y)$,
$b = (E(X) - E(Y))/a$, and f_Z and F_Z are the PDF and CDF of $N(0, 1)$, respectively.

In our example, $E(X) = 5$, $E(Y) = 5.25$, $V(X) = 0.25^2$, $V(Y) = 1$, $C(X, Y) = 0$. Thus, $E(W) \approx 5.5483$, and $V(W) \approx 0.4576$. The mean and variance of the MOP approximation of W are accurate to three decimal places. Unfortunately, the reasoning behind this computation of the marginal of W is not included in inference in Bayesian networks.

Another solution to computing the marginal of W using Bayesian network inference is to convert the *Max* function to a differentiable function as follows. $Max\{X, Y\} = X$ if $X \geq Y$, and $= Y$ if $X < Y$. We include a discrete variable A with two states, a and na, where a indicates that $X \geq Y$, and make it a parent of W. The revised Bayesian network is shown in Figure 14.

Figure 14. The revised Bayesian network for the max deterministic function



Starting with the Bayesian network in Figure 14, the marginal of W can be computed using the extended Shenoy-Shafer architecture described in Section 3. We start with mixed potentials as follows.

$$\mu_X(x) = (1, \xi(x)); \tag{4.6}$$

$$\mu_Y(y) = (1, \psi(y)); \tag{4.7}$$

$$\mu_A(a, x, y) = (H(x - y), 1), \mu_A(na, x, y) = (1 - H(x - y), 1); \text{ and} \tag{4.8}$$

$$\mu_W(a, x, y, w) = (1, \delta(w - x)), \mu_W(na, x, y, w) = (1, \delta(w - y)). \tag{4.9}$$

In (4.8), $H(.)$ is the Heaviside function such that $H(x) = 1$ if $x \geq 0$, and $= 0$ otherwise. The Heaviside function is a MOP function.

To find the marginal for W, we sequentially delete X, Y, and A. To delete X, first we combine μ_X, μ_A, and μ_W, and then marginalize X from the combination.

$$(\mu_X \otimes \mu_A \otimes \mu_W)(a, x, y, w) = (H(x - y), \xi(x)\, \delta(w - x)),$$
$$(\mu_X \otimes \mu_A \otimes \mu_W)(na, x, y, w) = (1 - H(x - y), \xi(x)\, \delta(w - y));$$
$$(\mu_X \otimes \mu_A \otimes \mu_W)^{-X}(a, y, w) = (1, \int H(x - y)\, \xi(x)\, \delta(w - x)\, dx) = (1, H(w - y)\, \xi(w)),$$
$$(\mu_X \otimes \mu_A \otimes \mu_W)^{-X}(na, y, w) = (1, \int (1 - H(x - y))\, \xi(x)\, \delta(w - y)\, dx)$$
$$= (1, \delta(w - y) \int (1 - H(x - y))\, \xi(x)\, dx) = (1, \delta(w - y)\, \theta(y)), \text{ where}$$
$$\theta(y) = \int (1 - H(x - y))\, \xi(x)\, dx.$$

Next, we delete Y. To do so, we combine $(\mu_X \otimes \mu_A \otimes \mu_W)^{-X}$ and μ_Y, and then marginalize Y.

$$((\mu_X \otimes \mu_A \otimes \mu_W)^{-X} \otimes \mu_Y)(a, y, w) = (1, H(w - y)\, \xi(w)\, \psi(y)),$$
$$((\mu_X \otimes \mu_A \otimes \mu_W)^{-X} \otimes \mu_Y)(na, y, w) = (1, \delta(w - y)\, \theta(y)\, \psi(y));$$
$$((\mu_X \otimes \mu_A \otimes \mu_W)^{-X} \otimes \mu_Y)^{-Y}(a, w) = (1, \xi(w) \int H(w - y)\, \psi(y)\, dy) = (1, \xi(w)\, \rho(w)), \text{ where}$$
$$\rho(w) = \int H(w - y)\, \psi(y)\, dy,$$
$$((\mu_X \otimes \mu_A \otimes \mu_W)^{-X} \otimes \mu_Y)^{-Y}(na, w) = (1, \theta(w)\, \psi(w)),$$

Finally, we delete A by marginalizing A out of $((\mu_X \otimes \mu_A \otimes \mu_W)^{-X} \otimes \mu_Y)^{-Y}$.

$(((\mu_X \otimes \mu_A \otimes \mu_W)^{-X} \otimes \mu_Y)^{-Y})^{-A}(w) = (1, \xi(w)\ \rho(w) + \theta(w)\ \psi(w)) = (1, \omega(w))$, where
$\omega(w) = \xi(w)\ \rho(w) + \theta(w)\ \psi(w)$.

$\omega(w)$ is a MOP approximation of $f_W(w)$. Notice that

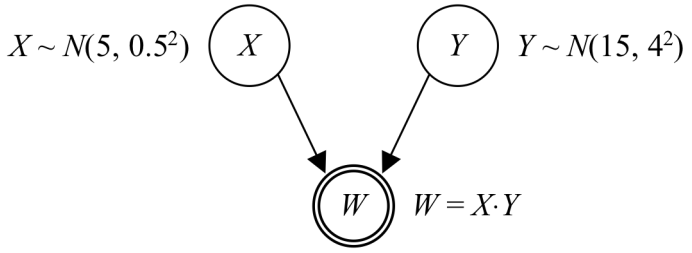
$$\rho(w) = \int H(w - y)\ \psi(y)\ dy = P(Y \leq w) = F_Y(w), \text{ and}$$
$$\theta(w) = \int (1 - H(x - w))\ \xi(x)\ dx = 1 - P(X > w) = F_X(w),$$

and therefore, $\omega(w) = \xi(w)\ \rho(w) + \theta(w)\ \psi(w)$ is a MOP approximation of $f_X(w)\ F_Y(w) + F_X(w)\ f_Y(w)$. We get exactly the same results as those obtained by order statistics, but using Bayesian network inference.

4.3 Example 3: The Product Deterministic Function

The main point of this example is to demonstrate that not all deterministic functions can be handled by MOP approximations. Consider a BN as shown in Figure 15.

Figure 15. A BN containing a deterministic variable with the product function.



We approximate the PDF of X and Y by MOP approximations as follows.

$$\xi(x) = 2\ \varphi_{6p}(2(x - 5));$$
$$\psi(y) = (1/4)\ \varphi_{6p}((y - 15)/4),$$

where $\varphi_{6p}(z)$ is the 6-piece, 3-degree MOP approximation of the PDF of $N(0, 1)$ on the interval $(-3, 3)$ described in (2.4). Notice that $\xi(x)$ and $\psi(y)$ are piecewise functions that are non-zero on the intervals (3.5, 6.5) and (3, 27), respectively. The deterministic conditional of W is represented by the Dirac potential $\omega(x, y, w) = \delta(w - x\ y)$. To find the marginal distribution of W, first we marginalize X, resulting in the potential

$$(\xi \otimes \omega)^{-X}(w, y) = \int \xi(x)\ \delta(w - x\ y)\ dx = (1/|y|) \int \xi(x)\ \delta(x - w/y)\ dx = (1/|y|)\ \xi(w/y).$$

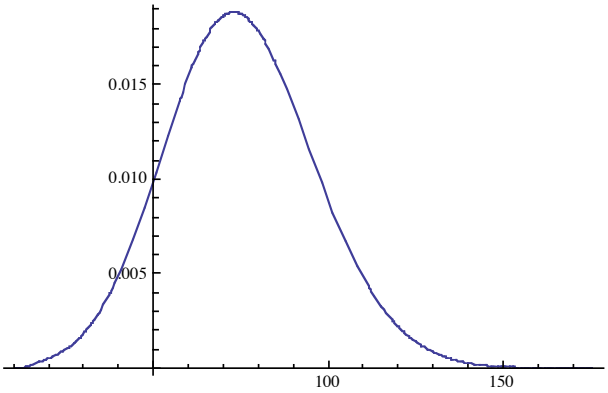
Next we marginalize Y. This results in the potential ω' for W as follows.

$$\omega'(w) = \int (1/|y|)\ \xi(w/y)\ \psi(y)\ dy \qquad (4.10)$$

Notice that the integrand $(1/y)\ \xi(w/y)\ \psi(y)$ is not a MOP function since neither $1/|y|$ nor $\xi(w/y)$ are MOP functions. Nevertheless, given the nature of the integrand, it is integrable.

The integral in (4.10) was evaluated using Mathematica®. The evaluation took 608 seconds on a desktop PC. The resulting function, $\omega'(w)$, is not a MOP function, but remains integrable. A graph of $\omega'(w)$ is shown in Figure 16.

Figure 16. A graph of $\omega'(w)$



It is confirmed that $\int \omega'(w)\, dw = 1$. The mean and variance of ω' were computed as 75.0 and 448.293, respectively. These two computations took 248 and 253 seconds, respectively. The exact mean of W is $E(X)\, E(Y) = 5 \cdot 15 = 75$. We do not know the exact variance of W. Using Monte Carlo simulation, the variance of W is estimated to be 460.42 with a 95% confidence interval of (459.83, 461.02).

If a deterministic variable with the product function is the main variable of interest, then using MOP functions is viable. If it is not the main variable of interest, then using MOP functions is not.

In summary, in this section, we have described three examples of hybrid Bayesian networks with deterministic variables. In the first example, we have a deterministic variable with the quotient function. The point of this example is to demonstrate that MOP functions are closed for quotient deterministic functions. In the second example, we have a hybrid Bayesian network with a deterministic variable with a non-differentiable function, $W = max\{X, Y\}$. This deterministic function can be converted to a linear differentiable function by the addition of a discrete indicator variable that tells us whether $X \geq Y$ or $X < Y$. This allows us to use MOPs for inference. In the third example, we have a hybrid Bayesian network with a deterministic variable $W = X \cdot Y$. MOPs are not closed under the transformations needed to find the marginal for W. However, even though the marginal for W is not a MOP, it remains integrable. Thus, if W is the variable of interest, then we can use MOPs for finding the marginal of W. If W is just an intermediate variable that is a parent of other variables of interest, then since the marginal of W is not a MOP, continuing with a non-MOP function may not be viable.

5 Summary and Discussion

The biggest problem associated with inference in hybrid BNs is the integration involved in marginalization of continuous variables. As a remedy, we have proposed MOP approximations for PDFs in the same spirit as MTE approximations [Moral *et al*. 2001]. Like MTE functions, MOP functions are easy to integrate, and are closed under combination and marginalization. This allows propagation of MOP potentials using the extended Shenoy-Shafer architecture [Shenoy and West 2009a].

MOP approximations have several advantages over MTE approximations of PDFs. First, they are easy to find using the Taylor series expansion of differentiable functions. Second, finding MOP approximations of multi-dimensional conditional PDFs is also relatively straightforward using the multi-dimensional Taylor series expansion. Third, MOP approximations are closed for a larger family of deterministic functions including the quotient functions. Beyond these observations, a formal empirical comparison of MOP vs. MTE approximations is an issue that needs further study.

Some issues associated with MOP approximations that need to be investigated are as follows. There is a tradeoff between the number of pieces and the degree of the polynomial. More pieces mean smaller intervals and consequently smaller degrees. Assuming the goal is to find marginals most efficiently, what is the optimal # pieces/degrees?

Another challenge is to describe the effect of pieces/terms on the errors in the moments of marginals. It appears that most of the errors in the moments are caused by truncating the domain of variables to some finite intervals. Thus, it may be possible to decide on what intervals should be used if we wish to compute marginals within some prescribed error bounds for the moments of the marginal of variable of interest.

High degree MOP approximations lead to very small coefficients that need to be kept in rational form. This may decrease the efficiency of computation, and may limit the size of BN models that can be solved. One solution here is to use more pieces, which lowers the degrees of the MOP approximations.

MOP approximations are not closed for many classes of deterministic functions such as products and exponentiation. If we can expand the class of MOP functions to include positive and negative rational exponents and maintain the properties of MOP functions —easily integrable, closed under combination and marginalization—then we can solve hybrid BNs with a larger class of deterministic functions.

Acknowledgments

We are grateful for some constructive comments from two anonymous UAI-09 reviewers, comments received after presentation at WUPES-09, and extensive comments received from three anonymous reviewers of *International Journal of Approximate Reasoning*. A condensed version of the paper appeared as [Shenoy and West 2009b].

References

Cinicioglu, E. N. and P. P. Shenoy (2009), "Arc reversals in hybrid Bayesian networks with deterministic variables," *International Journal of Approximate Reasoning*, **50**(5): 763–777.

Clark, C. E. (1961), "The greatest of a finite set of random variables," *Operations Research*, **9**(2): 145–162.

Cobb, B. R. and P. P. Shenoy (2005a), "Hybrid Bayesian networks with linear deterministic variables," *Uncertainty in Artificial Intelligence: Proceedings of the Twenty-First Conference* (UAI-05), F. Bacchus and T. Jaakkola (eds.), Corvallis, OR, AUAI Publishers: 136–144.

Cobb, B. R. and P. P. Shenoy (2005b), "Nonlinear deterministic relationships in Bayesian networks," in *Symbolic and Quantitative Approaches to Reasoning with Uncertainty* (ECSQARU-05), L. Godo (ed.), Berlin, Springer-Verlag, Lecture Notes in Artificial Intelligence, **3571**: 27–38.

Cobb, B. R. and P. P. Shenoy (2006), "Inference in hybrid Bayesian networks with mixtures of truncated exponentials," *International Journal of Approximate Reasoning*, **41**(3), 257–286.

Cobb, B. R., P. P. Shenoy, and R. Rumi (2006), "Approximating probability density functions in hybrid Bayesian networks with mixtures of truncated exponentials," *Statistics and Computing*, **16**(3): 293–308.

Curds, R. M. (1997), "Propagation techniques in probabilistic expert systems," PhD Thesis, Department of Statistical Science, University College London.

De Boor, C. (2001), *A Practical Guide to Splines*, revised ed., Springer-Verlag, New York.

Dirac, P. A. M. (1927), "The physical interpretation of the quantum dynamics," *Proceedings of the Royal Society of London*, *series A*, **113**(765): 621–641.

Dirac, P. A. M. (1958), *The Principles of Quantum Mechanics*, 4th ed., Oxford University Press, London.

Hoskins, R. F. (1979), *Generalised Functions*, Ellis Horwood, Chichester.

Howard, R. A. and J. E. Matheson (1984), "Influence diagrams," in *The Principles and Applications of Decision Analysis*, R. A. Howard and J. E. Matheson (eds.), Menlo Park, CA, Strategic Decisions Group, **2**: 719–762.

Jensen, F. V., S. L. Lauritzen, and K. G. Olesen (1990), "Bayesian updating in causal probabilistic networks by local computation," *Computational Statistics Quarterly*, **4**: 269–282.

Jensen, F., F. V. Jensen, and S. L. Dittmer (1994), "From influence diagrams to junction trees," in *Uncertainty in Artificial Intelligence*: *Proceedings of the Tenth Conference* (UAI-94), R. L. Mantaras and D. Poole (eds.), San Francisco, CA, Morgan Kaufmann: 367–373.

Kanwal, R. P. (1998), *Generalized Functions: Theory and Technique*, 2nd ed., Birkhäuser, Boston.

Khuri, A. I. (2004), "Applications of Dirac's delta function in statistics," *International Journal of Mathematical Education in Science and Technology*, **32**(2): 185–195.

Langseth, H., T. D. Nielsen, R. Rumi, and A. Salmeron (2010), "Parameter estimation and model selection for mixtures of truncated exponentials," *International Journal of Approximate Reasoning*, **51**(5), 485–498.

Lauritzen, S. L. and D. J. Spiegelhalter (1988), "Local computations with probabilities on graphical structures and their application to expert systems (with discussion)," *Journal of Royal Statistical Society, Series B*, **50**(2): 157–224.

Lauritzen, S. L. and F. Jensen (2001), "Stable local computation with conditional Gaussian distributions," *Statistics and Computing*, **11**: 191–203.

Lerner, U. N. (2002), "Hybrid Bayesian networks for reasoning about complex systems," PhD thesis, Department of Computer Science, Stanford University.

Lerner, U. N. and R. Parr (2001), "Inference in hybrid networks: Theoretical limits and practical algorithms," in *Uncertainty in Artificial Intelligence*: *Proceedings of the Seventeenth Conference* (UAI-01), J. Breese and D. Koller (eds.), Morgan Kaufmann, San Francisco, CA: 310–318.

Madsen, A. L. (2008), "Belief update in CLG Bayesian networks with lazy propagation," *International Journal of Approximate Reasoning*, **49**(2), 503–521.

Moral, S., R. Rumi, A. Salmeron (2001), "Mixtures of truncated exponentials in hybrid Bayesian networks," in *Symbolic and Quantitative Approaches to Reasoning under Uncertainty*

(ECSQARU-2001), P. Besnard and S. Benferhat (eds.), Berlin, Springer-Verlag, Lecture Notes in Artificial Intelligence, **2143**: 156–167.

Murphy, K. (1999), "A variational approximation for Bayesian networks with discrete and continuous latent variables," in *Uncertainty in Artificial Intelligence: Proceedings of the Fifteenth Conference* (UAI-99), K. B. Laskey and H. Prade (eds.), Morgan Kaufmann, San Francisco, CA: 457–466.

Olmsted, S. M. (1983), "On representing and solving decision problems," PhD dissertation, Department of Engineering-Economic Systems, Stanford University, Stanford, CA.

Pearl, J. (1986), "Fusion, propagation and structuring in belief networks," *Artificial Intelligence*, **29**: 241–288.

Romero, V., R. Rumi, and A. Salmeron (2006), "Learning hybrid Bayesian networks using mixtures of truncated exponentials," *International Journal of Approximate Reasoning*, **42**(1–2), 54–68.

Rumi, R. and A. Salmeron (2007), "Approximate probability propagation with mixtures of truncated exponentials," *International Journal of Approximate Reasoning*, **45**(2), 191–200.

Saichev, A. I. and W. A. Woyczyński (1997), *Distributions in the Physical and Engineering Sciences*, **1**, Birkhäuser, Boston.

Schoenberg, I. J. (1946), "Contributions to the problem of approximation of equidistant data by analytic functions," *Quarterly of Applied Mathematics*, **4**, 45–99, 112–141.

Schumaker, L. (2007), *Spline Functions: Basic Theory*, 3rd ed., Cambridge Univ. Press, New York.

Shachter, R. D. (1986), "Evaluating influence diagrams," *Operations Research*, **34**(6): 871-882.

Shenoy, P. P. (1992), "Valuation-based systems for Bayesian decision analysis," *Operations Research*, **40**(3): 463–484.

Shenoy, P. P. (2006), "Inference in hybrid Bayesian networks using mixtures of Gaussians," in *Uncertainty in Artificial Intelligence: Proceedings of the Twenty-Second Conference* (UAI-06), R. Dechter and T. Richardson (eds.), Corvallis, OR, AUAI Press: 428–436.

Shenoy, P. P. and G. Shafer (1990), "Axioms for probability and belief-function propagation," in *Uncertainty in Artificial Intelligence* 4 (UAI-88), R. D. Shachter, T. S. Levitt, J. F. Lemmer and L. N. Kanal (eds.), Amsterdam, North-Holland**, 4**:169–198.

Shenoy, P. P. and J. C. West (2009a), "Inference in hybrid Bayesian networks with deterministic variables," in *Symbolic and Quantitative Approaches to Reasoning under Uncertainty*–10*th* *ECSQARU*, C. Sossai and G. Chemello (eds.), Berlin, Springer-Verlag, Lecture Notes in Artificial Intelligence, **5590**: 46–58.

Shenoy, P. P. and J. C. West (2009b), "Mixtures of polynomials in hybrid Bayesian networks with deterministic variables," in *Proceedings of the 8th Workshop on Uncertainty Processing* (WUPES-09), T. Kroupa and J. Vejnarová (eds.), Prague, University of Economics: 202–212.