PREPRINT

Helge Langseth, Thomas D. Nielsen,
Rafael Rumí, and Antonio Salmerón:

Mixtures of Truncated Basis Functions

Cite preprint as:
Technical Report No. 2011-01
Department of Computer Science,
Aalborg University

To appear in:
International Journal of Approximate Reasoning
DOI: 10.1016/j.ijar.2011.10.004

Mixtures of Truncated Basis Functions

Helge Langseth, Thomas D. Nielsen, Rafael Rumí, and Antonio Salmerón

October 25, 2011

Abstract

In this paper we propose a framework, called mixtures of truncated basis functions (MoTBFs), for representing general hybrid Bayesian networks. The proposed framework generalizes both the mixture of truncated exponentials (MTEs) framework and the mixture of polynomials (MoPs) framework. Similar to MTEs and MoPs, MoTBFs are defined so that the potentials are closed under combination and marginalization, which ensures that inference in MoTBF networks can be performed efficiently using the Shafer-Shenoy architecture.

Based on a generalized Fourier series approximation, we devise a method for efficiently approximating an arbitrary density function using the MoTBF framework. The translation method is more flexible than existing MTE or MoP-based methods, and it supports an online/anytime tradeoff between the accuracy and the complexity of the approximation. Experimental results show that the approximations obtained are either comparable or significantly better than the approximations obtained using existing methods.

1 Introduction

Domains involving both discrete and continuous variables represent a challenge to Bayesian networks (BNs). The main difficulty is to find a representation of the joint distribution of the continuous and discrete variables that supports an efficient implementation of the usual inference operations over Bayesian networks (like *restriction*, *combination*, and *marginalization*, which are found in junction tree-based algorithms for exact inference). If all variables in the domain are discrete, their distributions can be represented by tables of probability values. This representation is very favorable from an operational point of view, as all three operations can be performed efficiently on this data structure. Furthermore, the operations are *closed* for probability tables, meaning that all the operations required during the inference process will always result in data that can conveniently be stored in probability tables. Unfortunately, inference becomes more complex when the domain involves continuous variables. From an implementation point of view we have no guarantee that a single data structure can be used to represent all intermediate results, but a more fundamental problem is that the results may no longer be available analytically, making exact inference unobtainable.

There are a number of popular strategies to overcome this problem: Firstly, one may choose to carefully construct the model so that exact inference algorithms can be applied. Computationally, this requires that the joint distribution over the variables of the domain is from a distribution-class that is closed under combination and marginalization. One example is the so-called Conditional Linear Gaussian (CLG) model (Lauritzen, 1992), where the joint distribution of the continuous variables conditioned on the discrete variables is assumed to

be a multivariate Gaussian. This puts some restrictions on the topology of the network. For example, discrete variables can only have discrete variables as parents, and continuous parents are to be seen as partial regression coefficients for their children. A second approach for addressing the inference problem is to apply *approximate inference*, for example using the Gibbs sampler (Geman and Geman, 1984; Hrycej, 1990). Next, variational techniques (Jordan et al., 1999) have recently gained much attention from the research community. Finally, the approach we will follow in this paper is to "translate" the original model into an *approximate model*, for which exact inference algorithms can be applied. The most common way of making this translation is by performing a *discretization* of the continuous variables (Friedman and Goldszmidt, 1996; Kozlov and Koller, 1997). Mathematically, this amounts to approximating the density function of every continuous variable by a step-function, which in turn implies that we can represent any conditional or joint distribution using a table. Unfortunately, discretization of variables can lead to a dramatic loss in precision, which is one of the reasons why other approaches have received much attention over the last few years. One of these alternatives is the *mixtures of truncated exponentials* (MTE) framework (Moral et al., 2001). This model can be seen as a generalization of discretization, since the density function is approximated by a sum of truncated exponential functions instead of a constant. The MTE framework therefore achieves more accurate approximations of the true density than standard discretization, even when using a smaller number of intervals and parameters. One of the advantages of this representation is that MTE distributions allow discrete and continuous variables to be treated in a uniform fashion, and since the family of MTEs is closed under combination and marginalization, inference in an MTE network can be performed efficiently using the Shafer-Shenoy architecture (Shafer and Shenoy, 1990; Cobb and Shenoy, 2006).

Cobb et al. (2006) empirically showed that many univariate distributions can be approximated accurately by means of an MTE distribution, and they argue that this makes the MTE framework an attractive general-purpose framework for Bayesian network modelling. Their main contribution is a library of ready-made MTE approximations for standard univariate densities. The MTE approximations were chosen to fit the original distributions with some accuracy, but, unfortunately, the general procedure does not include methods for i) finding an approximation with the quality requirements set at run-time or ii) easily obtaining approximations of other densities. Furthermore, the work by Cobb et al. (2006) has not been extended to handle *conditional* distributions nor distributions over more than one variable.

In this paper we propose a new procedure for translating a hybrid Bayesian network into an approximate model that supports exact inference. The theory we develop has a broader applicability than just MTEs. The procedure is therefore embedded in a framework called *Mixtures of Truncated Basis Functions* (MoTBFs) that includes MTEs and the recently proposed *Mixture of Polynomials* (MoP) framework (Shenoy and West, 2011) as special cases. We show how any hybrid BN can be efficiently translated into an MoTBF model. We propose an anytime algorithm that iteratively refines the approximation, choosing the refinements greedily, and which is able to meet arbitrary requirements regarding quality of the approximation. The paper is concluded by a small experimental study, where we investigate the properties of the approximations, and as a special case, show how our new approach outperforms the library-results obtained by Cobb et al. (2006).

2 The MoTBF-framework

In this section we introduce our mixture of truncated basis functions framework for representing hybrid Bayesian networks. Before doing so, we will briefly consider the starting-point for the framework, namely the mixtures of truncated exponentials framework (Moral et al., 2001) and the mixture of polynomials framework (Shenoy and West, 2011).

2.1 Background

A mixture of truncated exponential function can be seen as a generalized form of discretization, but instead of using a constant as approximation within each interval we use a linear combination of exponential functions. More formally, an MTE potential is defined as follows:

Definition 1. *Let $\boldsymbol{X}$ be a mixed n-dimensional random vector. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$ and $\mathbf{Z} = (Z_1, \ldots, Z_c)$ be the discrete and continuous parts of $\boldsymbol{X}$ respectively, with $c + d = n$. We say that a function $f : \Omega_{\boldsymbol{X}} \mapsto \mathbb{R}_0^+$ is a* mixture of truncated exponentials potential (MTE potential) *if one of the following two conditions holds:*

1. *f can be written as*

$$f(\boldsymbol{x}) = f(\boldsymbol{y}, \boldsymbol{z}) = a_{0,\boldsymbol{y}} + \sum_{i=1}^{k} a_{i,\boldsymbol{y}} \exp\left\{ \sum_{j=1}^{c} b_{i,\boldsymbol{y}}^{(j)} z_j \right\}, \tag{1}$$

   for all $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}$, where $a_{i,\boldsymbol{y}}$, $i = 0, \ldots, k$ and $b_{i,\boldsymbol{y}}^{(j)}$, $i = 1, \ldots, k$, $j = 1, \ldots, c$ are real numbers.

2. *There is a partition $\mathcal{I}_1, \ldots, \mathcal{I}_m$ of $\Omega_{\boldsymbol{X}}$ for which the domain of the continuous variables, $\Omega_{\mathbf{Z}}$, is divided into hyper-cubes and such that f is defined as*

$$f(\boldsymbol{x}) = f_i(\boldsymbol{x}) \quad \text{if} \quad \boldsymbol{x} \in \mathcal{I}_i \ ,$$

   where each f_i, $i = 1, \ldots, m$ can be written in the form of Equation (1).

An MTE potential is said to be a density *if $\sum_{\boldsymbol{y} \in \Omega_{\boldsymbol{Y}}} \int_{\Omega_{\boldsymbol{z}}} \phi(\boldsymbol{y}, \boldsymbol{z}) d\boldsymbol{z} = 1$.*

From the definition of an MTE potential, we see that the class of MTE functions is closed under combination and marginalization. Thus, it supports exact inference using the Shafer-Shenoy propagation architecture. Unfortunately, MTEs are not closed under division, which also implies that the specification of conditional MTE densities may be problematic. To be more specific, consider an MTE potential $\phi(\boldsymbol{z}_1, \boldsymbol{z}_2)$. In order for ϕ to be a conditional density for $\boldsymbol{z}_1$ given $\boldsymbol{z}_2$ we should have that $\int_{\boldsymbol{z}_1} \phi(\boldsymbol{z}_1, \boldsymbol{z}_2) d\boldsymbol{z}_1 = 1$ for each $\boldsymbol{z}_2 \in \Omega_{\boldsymbol{Z}_2}$. When fixing one of the elements of $\boldsymbol{z}_2$, which we denote z, this requirement corresponds to the constraint

$$\begin{aligned} \frac{\partial}{\partial z} \int_{\boldsymbol{z}_1} \phi(\boldsymbol{z}_1, \boldsymbol{z}_2) \, d\boldsymbol{z}_1 &= \frac{\partial}{\partial z} \int_{\boldsymbol{z}_1} a_0 + \sum_{i=1}^{k} a_i \exp(\boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{z}_1 + \boldsymbol{c}_i^{\mathrm{T}} \boldsymbol{z}_2) \, d\boldsymbol{z}_1 \\ &= \sum_{i=1}^{k} a_i c_i^{(z)} \exp(\boldsymbol{c}_i^{\mathrm{T}} \boldsymbol{z}_2) \int_{\boldsymbol{z}_1} \exp(\boldsymbol{b}_i^{\mathrm{T}} \boldsymbol{z}_1) \, d\boldsymbol{z}_1 = 0. \end{aligned}$$

Thus, we have uncountable many constraints, but only $\mathcal{O}(k)$ parameters to satisfy the constraints. For this to hold we set $\boldsymbol{c}_i = \mathbf{0}$, which means that the conditioning variables only affect the density through the hyper-cubes over which the MTE is defined. Thus, for two sets $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ of continuous variables with $\Omega_{\boldsymbol{Z}_2}$ partitioned into the hyper-cubes $\mathcal{I}_1, \ldots, \mathcal{I}_m$, we define a conditional MTE density $f(\boldsymbol{z}_1|\boldsymbol{z}_2)$ as (for ease of presentation we disregard all discrete variables and possible partitionings of $\Omega_{\boldsymbol{Z}_1}$):

$$f(\boldsymbol{z}_1|\boldsymbol{z}_2) = a_{0,\ell} + \sum_{i=1}^{k} a_{i,\ell} \exp(\boldsymbol{b}_{i,\ell}^{\mathrm{T}} \boldsymbol{z}_1), \tag{2}$$

for $\boldsymbol{z}_2 \in \mathcal{I}_\ell$.

Example 1. *The following is an example (Fernández et al., 2010) of a conditional MTE density following the definition above. Observe that the conditioning variable Z_2 only influences the conditional density through the hyper-cubes.*

$$f(z_1|z_2) = \begin{cases} 1.26 - 1.15e^{0.006 z_1} & \text{if } 0 \leq z_1 < 13,\ 0.4 \leq z_2 < 5; \\ 1.18 - 1.16e^{0.0002 z_1} & \text{if } 13 \leq z_1 < 43,\ 0.4 \leq z_2 < 5; \\ 0.07 - 0.03e^{-0.4 z_1} + 0.0001e^{0.0004 z_1} & \text{if } 0 \leq z_1 < 5,\ 5 \leq z_2 < 19; \\ -0.99 + 1.03e^{0.001 z_1} & \text{if } 5 \leq z_1 < 43,\ 5 \leq z_2 < 19. \end{cases}$$

Shenoy and West (2011) propose the mixture of polynomials potential as an alternative to the MTE potential. For an MoP potential, the core function is a *polynomial* whereas the MTE potential utilize an *exponential*. Hence, the univariate MoP potential for a continuous variable z equals $f(z) = a_{0,\ell} + \sum_{i=1}^{k} a_{i,\ell}\ z^i$ for $z \in \mathcal{I}_\ell$, and for a multivariate continuous vector $\boldsymbol{z} = (z_1, \ldots, z_c)^{\mathrm{T}}$, the potential takes on the form $f(\boldsymbol{z}) = \prod_{j=1}^{c} \left\{ a_{0,\ell}^{(j)} + \sum_{i=1}^{k} a_{i,\ell}^{(j)}\ z_j^i \right\}$ for $\boldsymbol{z} \in \mathcal{I}_\ell$.

One can easily translate any distribution into an MoP using a Taylor series expansion around a given point $\boldsymbol{z}_0$, for instance the mode of the distribution or the midpoint of a hyper cube (Shenoy and West, 2011). While appealing in its simplicity, the approximation strategy gives no guarantees about the quality of the approximation. In fact, if $\boldsymbol{z}$ is "far" from $\boldsymbol{z}_0$, the approximation may require many terms to obtain a reasonable quality. Furthermore, there is no guarantee that a translation defined using a Taylor series is strictly positive.

In contrast to MTEs, the MoP framework does not directly define conditional distributions. Instead, the MoP framework defines joint distributions as above, and calculates conditional distributions as fractions of joint distributions. For instance, the conditional distribution $f(z_1|z_2)$ will be based on the fraction $f(z_1, z_2)/f(z_2)$. As the set of polynomials is not closed under division, the fraction is not necessarily an MoP, and situation-specific approximations must be conducted to find an MoP-approximation to the conditional distribution.

2.2 The MoTBF model

When comparing the MTE and the MoP models, one can see that the potentials share the same structure but differ in the type of core function being used. Based on this observation we propose a generalization of these frameworks, where we instead of the exponential/polynomial functions use the abstract notation of *basis functions* $\psi(\cdot)$.

Definition 2. *Let $\boldsymbol{X}$ be a mixed n-dimensional random vector. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_d)$ and $\mathbf{Z} = (Z_1, \ldots, Z_c)$ be the discrete and continuous parts of $\boldsymbol{X}$, respectively, with $c + d = n$. Let $\Psi = \{\psi_i(\cdot)\}_{i=0}^{\infty}$ with $\psi_i : \mathbb{R} \to \mathbb{R}$ define a collection of real basis functions. We say that a function $\hat{f} : \Omega_{\mathbf{X}} \mapsto \mathbb{R}_0^+$ is a* mixture of truncated basis functions *(MoTBF) potential to level k wrt. Ψ if one of the following two conditions holds:*

1. *$\hat{f}$ can be written as*

$$\hat{f}(\boldsymbol{x}) = \hat{f}(\boldsymbol{y}, \boldsymbol{z}) = \sum_{i=0}^{k} \prod_{j=1}^{c} a_{i,\boldsymbol{y}}^{(j)} \psi_i(z_j), \tag{3}$$

   where $a_{i,\boldsymbol{y}}^{(j)}$ are real numbers.

2. *There is a partition $\mathcal{I}_1, \ldots, \mathcal{I}_m$ of $\Omega_{\boldsymbol{X}}$ for which the domain of the continuous variables, $\Omega_{\mathbf{Z}}$, is divided into hyper-cubes and such that f is defined as*

$$f(\boldsymbol{x}) = f_\ell(\boldsymbol{x}) \quad \text{if} \quad \boldsymbol{x} \in \mathcal{I}_\ell,$$

   where each f_ℓ, $\ell = 1, \ldots, m$ can be written in the form of Equation (3).

An MoTBF potential is said to be a density *if $\sum_{\boldsymbol{y} \in \Omega_{\boldsymbol{Y}}} \int_{\Omega_{\boldsymbol{Z}}} \hat{f}(\boldsymbol{y}, \boldsymbol{z}) d\boldsymbol{z} = 1$.*

As a direct generalization of the MTE framework, the MoTBF definition of a conditional distribution mirrors that of the MTEs. Thus, the influence a set of continuous parent variables $\boldsymbol{Z}$ have on their child variable X is encoded only through the partitioning of $\Omega_{\boldsymbol{Z}}$ into hyper-cubes, and not directly in the functional form of $\hat{f}_\ell(x|\boldsymbol{z})$ inside each hyper-cube $\mathcal{I}_\ell$ (confer also Equation (2)).

We will make some assumptions regarding the properties of the basis functions, which, as we shall discuss in Section 3.2, ensure that MoTBF approximations can be made arbitrarily good; these assumptions also help ensure that an MoTBF potential is closed under combination and marginalization. Let $\mathcal{Q}$ be the set of all linear combinations of the members of Ψ, i.e., the set of all functions of the type $\sum_{i=0}^{\infty} \alpha_i \psi_i$ for real constants α_i. Then, Ψ is said to define a set of *legal basis functions* if the following conditions hold:

1. ψ_0 is constant in its argument.
2. $f \in \mathcal{Q}$, $g \in \mathcal{Q} \implies (f \cdot g) \in \mathcal{Q}$.
3. For any pair of real numbers s and t, there exists a function $f \in \mathcal{Q}$ such that $f(s) \neq f(t)$.

In this paper we shall only consider legal sets of basis functions unless explicitly stated otherwise.

Example 2. *If we define $\psi_i(x) = x^i$ for $i = 0, 1, \ldots$, then $\mathcal{Q}$ corresponds to the set of polynomials, and $\{\psi_i\}_{i=0}^{\infty}$ thus trivially fulfills the requirements for being a legal set of basis functions. Now, Definition 2 equals the MoP model for univariate distributions. For multivariate distributions the MoP and MoTBF frameworks are slightly different, as the MoP framework only indirectly defines the conditional distributions that the MoTBF framework represents explicitly.*

Next, by choosing $\Psi(x) = \{1, \exp(-x), \exp(x), \exp(-2x), \exp(2x), \ldots\}$*, the MoTBF potential equals the MTE potential of Definition 1. Again, it is trivial to verify that the set of exponentials define a legal set of basis functions.*

Finally, let $\Psi = \{1, \log(x), \log(2x), \log(3x), \ldots\}$ *be defined for* $x > 0$*. This is* not *a legal set of basis functions, since Requirement 2 is not met.*

As we see from the example above, both MTE potentials and MoP potentials relate to MoTBF potentials, and the MoTBFs framework can therefore be seen as providing a unified framework for both MTEs and MoPs. Furthermore, since the basis functions are closed under combination and marginalization, then so are the MoTBF potentials. The MoTBF framework therefore also supports exact inference in hybrid domains using Shenoy-Shafer propagation.

3 Approximating univariate distributions using MoTBFs

As opposed to the locally bounded Taylor-series expansions used for making MoP approximations, we will in the following develop a generalized Fourier-series expansion for the class of MoTBFs. This expansion will provide a common framework for performing MTE and MoP approximations, and it will alleviate the two most important short-comings of the Taylor-series approach: the method will provide global error bounds for the MoTBF approximation, and it will ensure that the MoTBF approximation is in fact a probability density. The former property also implies that the approximations can be made arbitrarily tight, even if we do not split the domain of the variable into intervals. It should be noted that we will amend the generalized Fourier series expansion due to the requirement that the approximation should be a proper density (in particular, it should be non-negative).

3.1 The geometry of approximations

Before we describe the method for finding MoTBF approximations, we will start by introducing the required notation by first considering approximations in the real vector-space $\mathbb{R}^n$.

Assume we have a set of orthonormal basis vectors $\{\boldsymbol{e}_i\}$, where $i = 0, \ldots, n-1$. A set of basis vectors is *orthonormal* if each vector is of length one and all vectors are pairwise perpendicular. By letting $\langle \cdot, \cdot \rangle$ denote the inner product on $\mathbb{R}^n$, i.e., $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{y}$, we have that $\langle \boldsymbol{e}_i, \boldsymbol{e}_i \rangle = 1$ for $i = 0, \ldots, n-1$ and $\langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle = 0$ for $j \neq i$. Based on the inner product we define the norm $\|\cdot\|$ of a vector $\boldsymbol{x}$ as $\|\boldsymbol{x}\| = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$. If we want to approximate a vector $\boldsymbol{f}$ in the span of the first basis vector, i.e., generate the approximation $\hat{\boldsymbol{f}}_0 = \alpha_0 \, \boldsymbol{e}_0$, it is well known that α_0 should be chosen as the projection of $\boldsymbol{f}$ onto $\boldsymbol{e}_0$, i.e., $\hat{\boldsymbol{f}}_0 = \langle \boldsymbol{e}_0, \boldsymbol{f} \rangle \, \boldsymbol{e}_0$. This choice ensures that of all vectors of the form $\alpha_0 \, \boldsymbol{e}_0$, $\hat{\boldsymbol{f}}_0$ is the one closest to $\boldsymbol{f}$, where the distance is measured by $\|\boldsymbol{f} - \hat{\boldsymbol{f}}_0\| = \sqrt{\langle \boldsymbol{f} - \hat{\boldsymbol{f}}_0, \boldsymbol{f} - \hat{\boldsymbol{f}}_0 \rangle}$. Similarly, say we want to approximate $\boldsymbol{f}$ in the span of the first *two* basis vectors, $\hat{\boldsymbol{f}}_1 = \tilde{\alpha}_0 \, \boldsymbol{e}_0 + \tilde{\alpha}_1 \, \boldsymbol{e}_1$. Since $\langle \boldsymbol{e}_0, \boldsymbol{e}_1 \rangle = 0$ the projection along the first axis is not changed, $\tilde{\alpha}_0 = \alpha_0$, and we likewise choose $\tilde{\alpha}_1 = \langle \boldsymbol{e}_1, \boldsymbol{f} \rangle$ for the second basis vector. In general, the best projection using the first k basis vector of $\mathbb{R}^n$ will be $\hat{\boldsymbol{f}}_k = \sum_{i=0}^{k-1} \langle \boldsymbol{e}_i, \boldsymbol{f} \rangle \, \boldsymbol{e}_i$, and the distance (or the error) between $\hat{\boldsymbol{f}}_k$ and $\boldsymbol{f}$ is then $\|\boldsymbol{f} - \hat{\boldsymbol{f}}_k\| = \sqrt{\sum_{i=k}^{n-1} \langle \boldsymbol{f}, \boldsymbol{e}_i \rangle^2}$. More generally, if $\boldsymbol{h}_k = \sum_{i=0}^{k-1} \beta_i \cdot \boldsymbol{e}_i$ for some fixed $k \leq n$, then $\mathrm{error}(\boldsymbol{f}, \boldsymbol{h}_k) = \|\boldsymbol{f} - \hat{\boldsymbol{h}}_k\| = \sqrt{\sum_{i=0}^{k-1} (\beta_i - \alpha_i)^2 + \|\boldsymbol{f} - \boldsymbol{f}_k\|^2}$, where $\alpha_i = \langle \boldsymbol{e}_i, \boldsymbol{f} \rangle$.

3.2 Approximations of functions

Although the concepts above are defined for approximations in real vector spaces, many of them carry over to approximations in real function spaces. In this paper, we shall consider the space $L^2[a,b]$ of quadratically integrable real functions over the interval $\Omega = [a,b]$ with a and b being finite, i.e., functions where

$$\int_a^b f(x)^2\,dx < \infty.$$

For two functions $f(x)$ and $g(x)$ defined on $\Omega \subset \mathbb{R}$, we define the inner product as

$$\langle f, g\rangle = \int_\Omega f(x)\,g(x)\,dx,$$

which together with $L^2[a,b]$ constitute a Hilbert space (see, e.g., Kreyzig (1978)). Clearly, all bounded real functions on $\Omega \subset \mathbb{R}$ are quadratically integrable, which e.g. include the Gaussian function.

Given a set of orthonormal basis functions $\{\phi_k\}_{k=0}^\infty$ in $L^2[a,b]$, we can approximate $f \in L^2[a,b]$ using $\hat{f} = \sum_i \langle f, \phi_i\rangle \cdot \phi_i$. This approximation is also known as a *Generalized Fourier Series* approximation. The function $\hat{f}$ minimizes $\int_\Omega (f(x) - \hat{f}(x))^2\,dx$, and, in particular, by using trigonometric basis functions we obtain the standard Fourier series.

For a set of functions $\{\psi_k\}_{k=0}^\infty$ defined on $\Omega \subset \mathbb{R}$ that is not orthonormal, say $\{1, \exp(-x), \exp(x), \exp(-2x), \ldots\}$ we can e.g. use the Gram–Schmidt process to obtain orthonormal functions $\{\phi_k\}_{k=0}^\infty$ such that for all $i \geq 0$ we have $\phi_i = \sum_{j=0}^i \alpha_{j,i}\psi_i$ for some constants $\alpha_{j,i}$.

If the functions $\Psi = \{\psi_k\}_{k=0}^\infty$ are *dense* in the space of all quadratically integrable functions, the generalized Fourier approximation can be made *arbitrarily good*. It is well-known that this is the case for polynomials (Weierstrass, 1885), but it also holds for any set Ψ of legal basis functions (Stone, 1937). As a consequence, we can obtain approximations that are *arbitrarily good* even *without* splitting Ω into sub-intervals. Furthermore, as a set of legal basis functions is closed under combination and marginalization, the derived approximations support inference in the Shenoy-Shafer architecture.

Example 3. *Assume we want to approximate a Gaussian distribution function with expected value $\mu = 0$ and standard deviation $\sigma = 0.386$. The interval of interest is $\Omega = [-1,1]$, containing 99% of the probability mass. It is well known that the Legendre-polynomials are orthonormal on the interval $\Omega = [-1,1]$. The unnormalized Legendre polynomial of order m, $P_m(x)$, is defined by*

$$P_m(x) = \frac{1}{2^m \cdot m!}\,\frac{\mathrm{d}^m}{\mathrm{d}x^m}\left(x^2 - 1\right)^m,$$

for $m = 0, 1, \ldots$. See Figure 1 for the first four Legendre polynomials.

Figure 2 (a) shows the approximation using only one (constant) function, i.e., $\hat{f}(x) = \langle f, \phi_0\rangle\,\phi_0(x)$. The approximation procedure ensures that the probability mass of $\hat{f}$ is allocated correctly (see also Section 3.3), but the approximation is obviously poor. Part (b) approximates the Gaussian using both a constant term and a linear term. The contribution from the linear term vanishes, since the Gaussian pdf is an equal function whereas the linear function is odd, meaning that their product is odd, and the integral becomes zero. This is the case for all $\phi_{2j+1}(x)$, $j = 0, 1, 2, \ldots$.

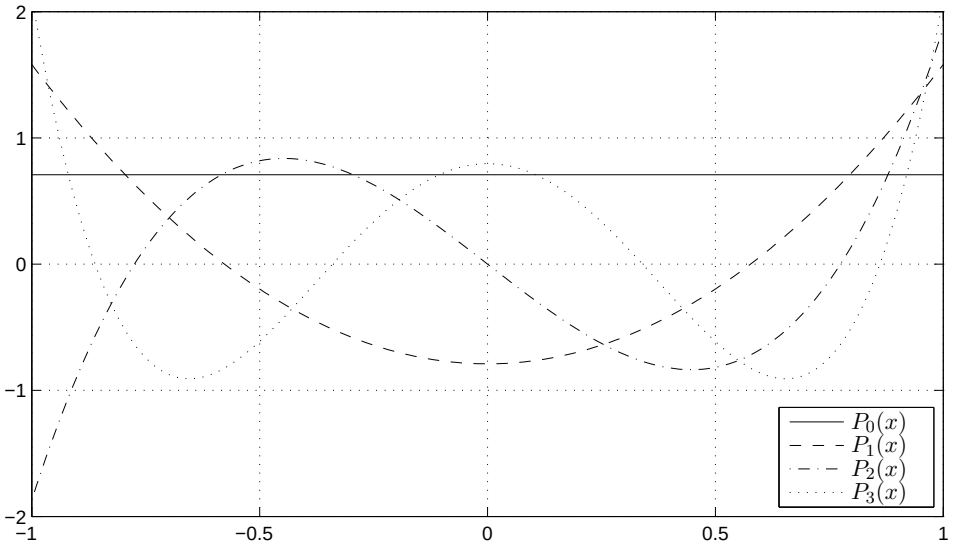


Figure 1: The first four Legendre polynomials on the interval $\Omega = [-1, 1]$.

Figure 2 (c) gives the approximation using three functions: $\hat{f}(x) = \sum_{i=0}^{2} \langle f, \phi_i \rangle \ \phi_i(x)$. *Note that since* $\langle \phi_0, \phi_2 \rangle = 0$ *the probability mass of* $\hat{f}$ *is unchanged. Finally, Figure 2 (d) shows an approximation using the five first* contributing *functions:*

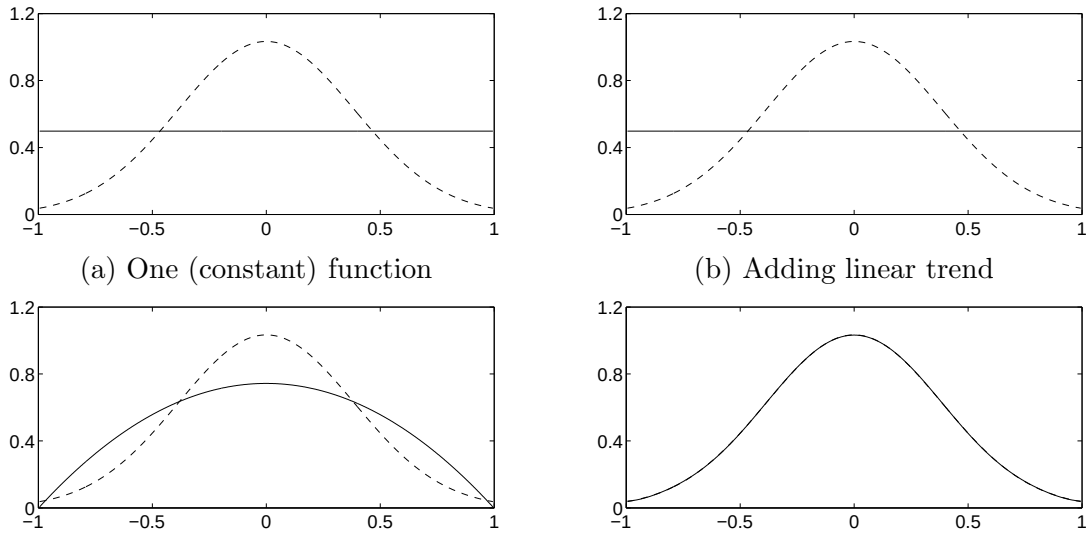
$$\hat{f}(x) = \sum_{i=0}^{4} \langle f, \phi_{2i} \rangle \ \phi_{2i}(x),$$

3.3 Ensuring that $\hat{f}$ is a density

So far, we have only chosen $\hat{f}$ to minimize $\int_\Omega (f(x) - \hat{f}(x))^2 \, dx$. Next, we will turn to the validity of the approximation. For $\hat{f}(x)$ defined on $x \in \Omega$ to be a density, we must have both that $\int_{x \in \Omega} \hat{f}(x) \, dx = 1$, and that $\hat{f}(x) \geq 0$, $\forall x \in \Omega$. The former constraint is easily verified. Remember that $\phi_0(x)$ is a constant, meaning that $\alpha_0 = \langle f, \phi_0 \rangle = \phi_0 \cdot \int_{x \in \Omega} f(x) \, dx$, and since $\|\phi_0\| = 1$ we have that $\int_{x \in \Omega} \alpha_0 \phi_0 \, dx = 1$. Furthermore, as $\langle \phi_0, \phi_j \rangle = 0$ for all $j \neq 0$, it follows that $\int_{x \in \Omega} \phi_j(x) \, dx = 0$, for all $j \neq 0$, and the term containing ϕ_0 is therefore the only term in the mixture that contributes to the probability mass over Ω.

For the latter requirement there are no equally simple results as $\hat{f}$ may be negative for some $x \in \Omega$. An example of this problem is given in Figure 3, where the left hand side of the figure gives an approximation to the χ^2_5 distribution using polynomials up to degree eight. However, when considering the approximation around $x = 0$ (the right panel) we see that the approximation attains negative values.

One solution to this problem is to minimize the error between $\hat{f}$ and f under the constraint that $\hat{f}(x) \geq 0, x \in \Omega$. From Section 3.1 we have that if $h_k(x) = \sum_{i=0}^{k-1} \alpha_i \phi_i(x)$ then the squared error of the approximation is $(\text{error}(f, h_k))^2 = \sum_{i=0}^{k-1} (\langle f, \phi_i \rangle - \alpha_i)^2 + \sum_{i=k}^{\infty} \langle f, \phi_i \rangle^2$. The latter sum depends only on the number of basis functions used in the approximation and not on



(a) One (constant) function

(b) Adding linear trend

(c) Adding quadratic term.

(d) Polynomials up to degree eight.

Figure 2: The figure shows four approximations to the normal distribution with mean value $\mu = 0$ and standard deviation $\sigma = 0.386$. The approximations are defined over the interval $\Omega = [-1, 1]$.

the chosen coefficients α_i. Hence, with a fixed number of basis functions we can focus on minimizing the first term in the error:

$$\begin{aligned} \textbf{Minimize} \quad & \sum_{i=0}^{k-1} \left(\langle f, \phi_i \rangle - \alpha_i \right)^2 \\ \textbf{Subject to} \quad & \sum_{i=0}^{k-1} \alpha_i \phi_i(x) \geq 0, x \in \Omega \\ & \alpha_0 = \langle \phi_0, f \rangle \end{aligned}$$

This is a *convex optimization problem*, and can be solved using algorithms for semi-definite programming (Vandenberghe and Boyd, 1996).

The optimization problem proposed above attempts to minimize the quadratic difference between the true distribution f and the approximation $\hat{f}$, and does not directly have a well-founded probabilistic interpretation. A more common measure of the distance from f to $\hat{f}$ is the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951).[1] Formally, the KL-divergence is defined as

$$D\left(f \parallel \hat{f} \right) = \int_{\Omega} f(\boldsymbol{x}) \log \left(\frac{f(\boldsymbol{x})}{\hat{f}(\boldsymbol{x})} \right) d\boldsymbol{x}. \tag{4}$$

There are many arguments for using this particular measurement for calculating the quality of the approximation, see Cover and Thomas (1991). One of them is the fact that the KL

[1] Strictly speaking the KL-divergence is not a distance measure, since it is neither symmetric nor does it satisfy the triangle inequality.

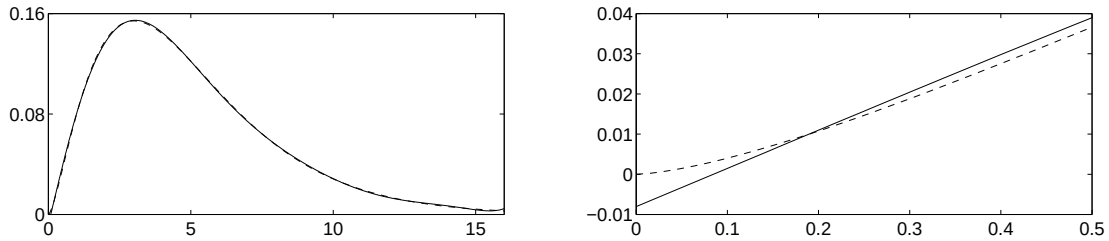


Figure 3: The χ^2 distribution with 5 degrees of freedoms approximated by nine polynomials. The actual pdf is given by the dashed line, and the approximation by the fat solid line. The full density is given on the left-hand side of the figure, whereas the right-hand side shows the results zoomed in at $x = 0$.

divergence bounds the maximum error in the assessed probability for any event E (Whittaker, 1990, Proposition 4.3.7):

$$\sup_E \left| \int_{\boldsymbol{x} \in E} f(\boldsymbol{x}) d\boldsymbol{x} - \int_{\boldsymbol{x} \in E} \hat{f}(\boldsymbol{x}) d\boldsymbol{x} \right| \leq \sqrt{\frac{1}{2} D\left(f(\boldsymbol{x}) \,\|\, \hat{f}(\boldsymbol{x}) \right)}.$$

Similar results for the maximal error of the estimated conditional distribution are derived by van Engelen (1997).

Another important property is that the KL divergence factorizes, so that for two random vectors $(\boldsymbol{X},\boldsymbol{Y})$, we have that $D\left(f(\boldsymbol{x},\boldsymbol{y}) \,\|\, \hat{f}(\boldsymbol{x},\boldsymbol{y}) \right) = D\left(f(\boldsymbol{x}) \,\|\, \hat{f}(\boldsymbol{x}) \right) + D\left(f(\boldsymbol{y}|\boldsymbol{x}) \,\|\, \hat{f}(\boldsymbol{y}|\boldsymbol{x}) \right)$, where the last term is defined as

$$D\left(f(\boldsymbol{y}|\boldsymbol{x}) \,\|\, \hat{f}(\boldsymbol{y}|\boldsymbol{x}) \right) = \mathbb{E}_{\boldsymbol{X}} \left[\int_{\boldsymbol{y}} f(\boldsymbol{y}|\boldsymbol{X}) \log \left(\frac{f(\boldsymbol{y}|\boldsymbol{X})}{\hat{f}(\boldsymbol{y}|\boldsymbol{X})} \right) \right].$$

This invites a divide-and-conquer type of strategy for finding good approximate distributions $\hat{f}$: One can look at one family at a time. From a computational point of view it is also interesting to note that the KL divergence factorizes according to the BN structure, and can therefore be calculated without expanding the full integral in Equation (4), see Cowell et al. (1999, Chapter 6). Finally, Zeevi and Meir (1997, Lemma 3.3) showed that the KL-divergence is related to the L^2-error:

$$D\left(f \,\|\, \hat{f} \right) \leq \frac{\left(\text{error}(f, \hat{f}) \right)^2}{\min_{x \in \Omega} \hat{f}(x)}. \tag{5}$$

We can use this relation when finding MoTBF approximations. That is, instead of minimizing the L^2-error (as in the optimization problem above), we can instead minimize the upper bound of the KL divergence defined in Equation 5. The updated optimization problem wrt. $(\boldsymbol{\alpha}, \xi)$ therefore becomes:

$$\begin{aligned} \textbf{Minimize } & \frac{1}{\xi} \cdot \sum_{i=0}^{k-1} \left(\langle f, \phi_i \rangle - \alpha_i \right)^2 \\ \textbf{Subject to } & \sum_{i=0}^{k-1} \alpha_i \phi_i(x) \geq \xi, x \in \Omega \\ & \alpha_0 = \langle \phi_0, f \rangle \end{aligned}$$

Note that the updated problem is still convex, and can still be solved using semi-definite programming.

4 Conditional Distributions

In the two sections below, we will consider translation methods for conditional distributions of discrete and continuous variable, respectively. For ease of exposition, we shall disregard any discrete parents of these variables, since any such variable will only serve to index the distribution in question and the results therefore generalize immediately.

4.1 Discrete Conditional Distributions

Consider a discrete variable Y with continuous parents $\boldsymbol{X}$. In order to translate the distribution $f(Y|\boldsymbol{X})$ (e.g. a logistic or probit function), we will again pose the problem as a convex optimization problem. As a simplifying assumption, we will for now assume that the discrete variable is binary in which case we can find a representation $\hat{f}(Y = q|\boldsymbol{X} = \boldsymbol{x})$ of $f(Y = q|\boldsymbol{X} = \boldsymbol{x})$ by solving the following optimization problem:

$$\begin{aligned}
\textbf{Minimize} \quad & \sum_{i=0}^{k-1} \left(\langle f, \phi_i \rangle - \alpha_i\right)^2 \\
\textbf{Subject to} \quad & \sum_{i=0}^{k-1} \alpha_i \phi_i(x) \geq 0, \boldsymbol{x} \in \Omega_{\boldsymbol{X}} \\
& \sum_{i=0}^{k-1} \alpha_i \phi_i(x) \leq 1, \boldsymbol{x} \in \Omega_{\boldsymbol{X}},
\end{aligned}$$

and using a suitable combination function for the continuous parent variables (e.g. with the logistic function we have a weighted linear combination of the parent variables). Observe that in this formulation of the problem we directly seek to minimize the L^2-error.

When the discrete variable is non-binary (for example associated with a soft-max function), the optimization procedure needs to simultaneously consider all the states of the variable in order to ensure that $\sum_{y \in \Omega_Y} \hat{f}(Y = y|\boldsymbol{X} = \boldsymbol{x}) = 1$, for all $\boldsymbol{x} \in \Omega_{\boldsymbol{X}}$.

Example 4. *Figure 4 shows two MoTBF approximations for the probit functions having weights 1 and 3, respectively. The approximations are defined for the interval* $[-2.58, 2.58]$ *using 5 and 10 polynomial basis functions. Observe that for a fixed set of basis functions, the quality of the approximation is very dependent on the weight being used.*

The quality and the result of the approximation will depend on the hyper-cube $\Omega_{\boldsymbol{X}}$ for which the MoTBF is specified. One approach is to define $\Omega_{\boldsymbol{X}}$ so that it covers the interval for which $\epsilon \leq P(Y = 0|\boldsymbol{X} = \boldsymbol{x}) \leq 1 - \epsilon$, for some $\epsilon > 0$. However, this approach does not take the density of $\boldsymbol{X}$ into account and, in particular, $\Omega_{\boldsymbol{X}}$ may therefore include regions with very low probability mass. Instead we define $\Omega_{\boldsymbol{X}}$ based on the density function for $\boldsymbol{X}$ s.t. $\Omega_{\boldsymbol{X}}$ covers a certain amount of the probability mass of $f(\boldsymbol{X})$. For instance, in the example above we have assumed that $X \sim N(0, 1)$ and define $\Omega_X = [a, b]$ such that $P(X \leq a) = 0.005$ and $P(X \leq b) = 0.995$.

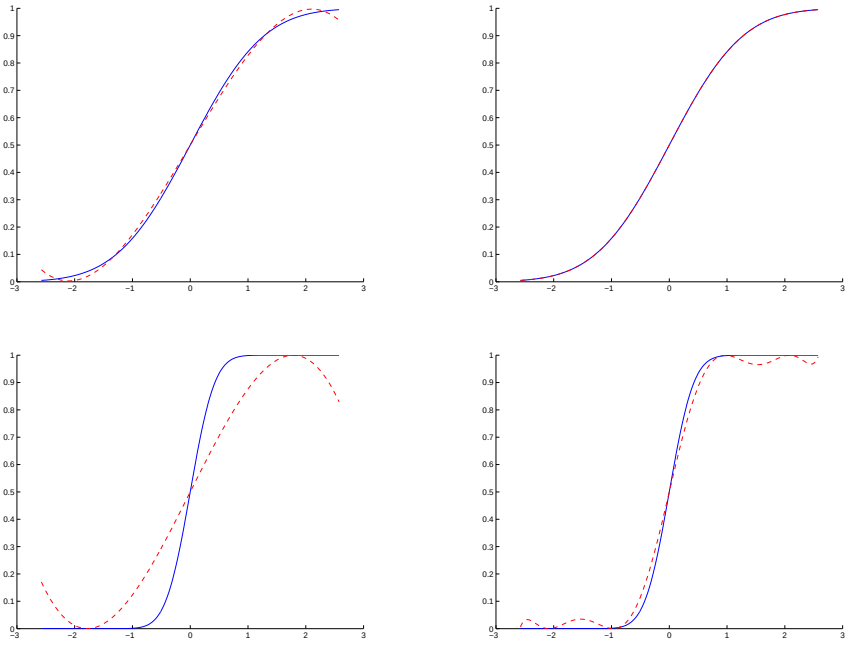


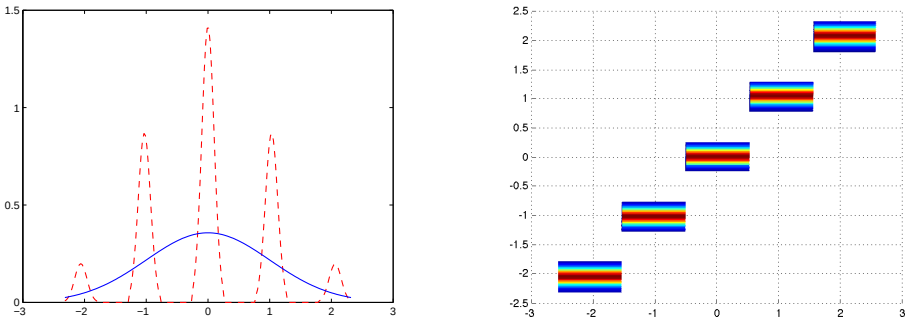
Figure 4: The first row shows MoTBF approximations for the probit function with weight 1 and offset 0 using 5 and 10 polynomial basis functions, respectively. The second row shows MoTBF approximations for the probit function with weight 3 and offset 0, also using 5 and 10 polynomial basis functions.

4.2 Continuous Conditional Distributions

In this section we shall consider methods for obtaining conditional MoTBFs densities for continuous variables, using the restricted type of MoTBFs where the conditioning variables only interact with the conditional density through the hyper-cubes and not the specific numerical values of the conditioning variables. With this constrained type of conditional MoTBFs, approximating a conditional density function reduces to finding a partitioning of the state space of the conditioning variables and, for each of these partitions, *i*) selecting the number of basis functions and *ii*) approximating the conditional density function by an MoTBF potential.

4.2.1 Finding an MoTBF approximation for a fixed partitioning

Consider a conditional density $f(y|\boldsymbol{x})$ and assume that we have found a partitioning $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_k$ of $\Omega_{\boldsymbol{X}}$. An immediate approach for finding an approximation for $f(y|\boldsymbol{x} \in \mathcal{I}_j)$ could be to simply approximate $f(y|\boldsymbol{x}_0)$ for some fixed $\boldsymbol{x}_0 \in \mathcal{I}_j$ (e.g. chosen as the midpoint or the mass-center in $\mathcal{I}_j$). Unfortunately, using this approach we will often underestimate the variance of $f(y|\boldsymbol{x} \in \boldsymbol{I}_j)$. To illustrate the effect, consider the conditional linear Gaussian model defined by $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(x, 0.1)$. Suppose that we look for an MoTBF approximation over the domain defined by $\Omega_X = [-2.58, 2.58]$ and $\Omega_Y = [-2.83, 2.83]$, and using a polynomial basis up to order 9. If $f(y|x)$ is approximated by making a five interval partitioning of Ω_X where we condition on the midpoints of these intervals, then we get the marginal density $\hat{f}(y)$ shown in Figure 5(a) and the conditional density $\hat{f}(y|x)$ shown in Figure 5(b).



(a) (b)

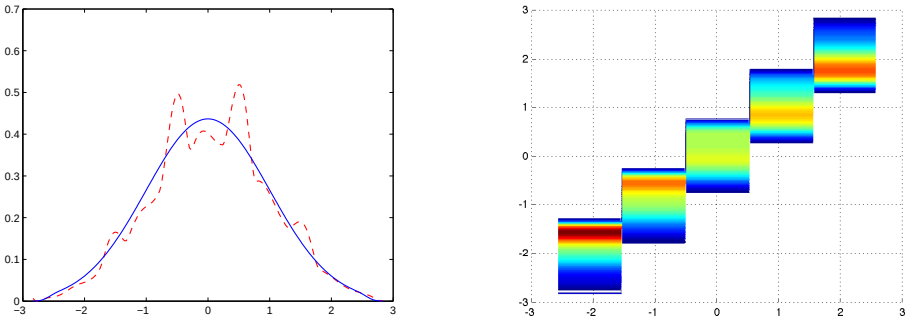
Figure 5: Figure (a) shows $f(y)$ and $\hat{f}(y)$ for the Gaussian model defined by $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(x, 0.1)$. The marginal density $\hat{f}(y)$ is obtained by approximating $f(x)$ and $f(y|x)$ based on a polynomial basis up to order 9 and by using the midpoints of a five interval partitioning of X when approximating $f(y|x)$. Figure (b) shows $\hat{f}(y|x)$.

Alternatively we can look for an MoTBFs representation of $f(y|\boldsymbol{x} \in \mathcal{I}_j)$:

$$\hat{f}(y|\boldsymbol{x}) \sim f(y|\boldsymbol{x} \in \mathcal{I}_j) = \int_{\boldsymbol{x}} f(y|\boldsymbol{x}) f(\boldsymbol{x}|\boldsymbol{x} \in \mathcal{I}_j),$$

where the integral can be approximated by $\sum_{i=1}^{n} f(y|\boldsymbol{x}_i) f(\boldsymbol{x}_i|\boldsymbol{x}_i \in \mathcal{I}_j)$ with samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ from $\mathcal{I}_j$. Using this approach for approximating the linear Gaussian model described above,

we get the marginal density $\hat{f}(y)$ shown in Figure 6(a) and the conditional density $\hat{f}(y|x)$ shown in Figure 6(b). This is the path we will pursue in the following.



(a) (b)

Figure 6: Figure (a) shows $f(y)$ and $\hat{f}(y)$ for the Gaussian model defined by $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(x, 0.1)$. The marginal density $\hat{f}(y)$ is obtained by approximating $f(x)$ and $f(y|x)$ based on a polynomial basis up to order 9, and by approximating $f(y|x)$ by conditioning on each of the five partitions on Ω_X. Figure (b) shows $\hat{f}(y|x)$.

Having decided what the conditional MoTBF potential should approximate, the next step is to consider the trade-off between model complexity and model quality. In particular, we calculate this as the decrease in KL divergence between the true model and the MoTBF representation, when the MoTBF representation uses $k+1$ instead of k basis functions for a particular hyper-cube $\boldsymbol{x} \in \mathcal{I}_j$. Let $\hat{f}_\ell(y|\boldsymbol{x} \in \mathcal{I}_j)$ be the MoTBF representation of $f(y|\boldsymbol{x} \in \mathcal{I}_j)$ using ℓ basis functions. Then, after some pencil-pushing, we find that the reduction in KL divergence is given by

$$D\left(f \,\|\, \hat{f}_k\right) - D\left(f \,\|\, \hat{f}_{k+1}\right) \;=\; \int_y \int_{\boldsymbol{x} \in \mathcal{I}_j} f(\boldsymbol{x}, y)\, d\boldsymbol{x} \log \frac{\hat{f}_{k+1}(y|\boldsymbol{x} \in \mathcal{I}_j)}{\hat{f}_k(y|\boldsymbol{x} \in \mathcal{I}_j)}\, dy.$$

The calculation of the KL improvement involves a multivariate integral, which potentially can be costly to evaluate. We therefore approximate $f(y|\boldsymbol{x})$ by $\hat{f}_{k+1}(y|\boldsymbol{x} \in \mathcal{I}_j)$, and obtain

$$\begin{aligned} D\left(f \,\|\, \hat{f}_k\right) - D\left(f \,\|\, \hat{f}_{k+1}\right) &\approx \int_{\boldsymbol{x} \in \mathcal{I}_j} f(\boldsymbol{x})\, d\boldsymbol{x} \cdot \int_y \hat{f}_{k+1}(y|\boldsymbol{x} \in \mathcal{I}_j) \log \frac{\hat{f}_{k+1}(y|\boldsymbol{x} \in \mathcal{I}_j)}{\hat{f}_k(y|\boldsymbol{x} \in \mathcal{I}_j)}\, dy \\ &= P(\boldsymbol{x} \in \mathcal{I}_j) \cdot D\left(\hat{f}_{k+1} \,\|\, \hat{f}_k\right). \end{aligned}$$

Hence, the reduction in KL-divergence can be approximated by the KL divergence between the two MoTBF representations, weighted by the probability of the parents being in that particular hyper-cube.

4.2.2 Finding a partitioning of the conditioning variables

When approximating a conditional density we need to find a partitioning of the state space of the conditioning variables $\Omega_{\boldsymbol{X}}$. In this paper we will pursue a myopic strategy (detailed later), where we at each step consider the reduction in KL-divergence obtained by dividing one of the existing partitions.

Consider a density $f(y|\boldsymbol{x})$ and an MoTBF approximation using n partitions $\mathcal{I}_1, \ldots, \mathcal{I}_n$ of $\Omega_{\boldsymbol{X}}$. Assume now that we divide the partition $\mathcal{I}_k$ into $\mathcal{I}_k^0$ and $\mathcal{I}_k^1$, which results in a conditional MoTBF $\hat{f}'(y|\boldsymbol{x})$ where we approximate $f(y|\boldsymbol{x})$ with $\hat{f}'(y|\boldsymbol{x} \in \mathcal{I}_k^0)$ and $\hat{f}'(y|\boldsymbol{x} \in \mathcal{I}_k^1)$ when $\boldsymbol{x} \in \mathcal{I}_k$ and $\hat{f}(y|\boldsymbol{x}) = \hat{f}'(y|\boldsymbol{x})$ for all the other partitions. The reduction in KL-divergence is then given by

$$\begin{aligned} D\left(f \,\|\, \hat{f}\right) - D\left(f \,\|\, \hat{f}'\right) = \quad & \int_{\boldsymbol{x} \in \mathcal{I}_k} \int_y f(\boldsymbol{x}, y) \log \frac{f(y|\boldsymbol{x})}{\hat{f}(y|\boldsymbol{x} \in \mathcal{I}_k)} \, dy \, d\boldsymbol{x} - \\ & \left(\int_{\boldsymbol{x} \in \mathcal{I}_k^0} \int_y f(\boldsymbol{x}, y) \log \frac{f(y|\boldsymbol{x})}{\hat{f}'(y|\boldsymbol{x} \in \mathcal{I}_k^0)} \, dy \, d\boldsymbol{x} + \right. \\ & \left. \int_{\boldsymbol{x} \in \mathcal{I}_k^1} \int_y f(\boldsymbol{x}, y) \log \frac{f(y|\boldsymbol{x})}{\hat{f}'(y|\boldsymbol{x} \in \mathcal{I}_k^1)} \, dy \, d\boldsymbol{x} \right) \\ = \quad & \int_y \int_{\boldsymbol{x} \in \mathcal{I}_k^0} f(\boldsymbol{x}) f(y|\boldsymbol{x}) \, d\boldsymbol{x} \log \frac{\hat{f}'(y|\boldsymbol{x} \in \mathcal{I}_k^0)}{\hat{f}(y|\boldsymbol{x} \in \mathcal{I}_k)} \, dy + \\ & \int_y \int_{\boldsymbol{x} \in \mathcal{I}_k^1} f(\boldsymbol{x}) f(y|\boldsymbol{x}) \, d\boldsymbol{x} \log \frac{\hat{f}'(y|\boldsymbol{x} \in \mathcal{I}_k^1)}{\hat{f}(y|\boldsymbol{x} \in \mathcal{I}_k)} \, dy. \end{aligned}$$

As above we are confronted by multi-dimensional integration, and this time we approximate $f(y|\boldsymbol{x})$ by $\hat{f}'_0(y|\boldsymbol{x} \in \mathcal{I}_k^0)$ and $\hat{f}'_1(y|\boldsymbol{x} \in \mathcal{I}_k^1)$ on $\mathcal{I}_k^0$ and $\mathcal{I}_k^1$, respectively:

$$\begin{aligned} D\left(f \,\|\, \hat{f}\right) - D\left(f \,\|\, \hat{f}'\right) \approx \quad & \int_{\boldsymbol{x} \in \mathcal{I}_k^0} f(\boldsymbol{x}) \, d\boldsymbol{x} \cdot \int_y \hat{f}'(y|\boldsymbol{x} \in \mathcal{I}_k^0) \log \frac{\hat{f}'(y|\boldsymbol{x} \in \mathcal{I}_k^0)}{\hat{f}'(y|\boldsymbol{x} \in \mathcal{I}_k)} \, dy + \\ & \int_{\boldsymbol{x} \in \mathcal{I}_k^1} f(\boldsymbol{x}) \, d\boldsymbol{x} \cdot \int_y \hat{f}'(y|\boldsymbol{x} \in \mathcal{I}_k^1) \log \frac{\hat{f}'(y|\boldsymbol{x} \in \mathcal{I}_k^1)}{\hat{f}(y|\boldsymbol{x} \in \mathcal{I}_k)} \, dy \\ = \quad & P(\boldsymbol{x} \in \mathcal{I}_k^0) \cdot D(f', f|\mathcal{I}_k^0) + P(\boldsymbol{x} \in \mathcal{I}_k^1) \cdot D(f', f|\mathcal{I}_k^1). \end{aligned}$$

Thus, the reduction in KL-divergence can be estimated by calculating the KL divergences $D(f', f|\mathcal{I}_k^0)$ and $D(f', f|\mathcal{I}_k^1)$.

5 The Overall Algorithm

Based on the methods for finding MoTBF representations of univariate and conditional distributions, we can now describe a general algorithm for approximating an arbitrary hybrid Bayesian network with an MoTBF network. The MoTBF network is initialized with MoTBF potentials defined by a single basis function and with no split points. The algorithm then iteratively selects a local MoTBF potential $\hat{f}$ to refine, using a heuristic selection criterion based on an estimate of the immediate decrease in KL-divergence per additional parameter introduced in the model:

$$h(\hat{f}', f, \hat{f}) = \frac{D\left(f \,\|\, \hat{f}\right) - D\left(f \,\|\, \hat{f}'\right)}{\dim(\hat{f}') - \dim(\hat{f})},$$

where f is the true distribution and $\hat{f}'$ is the refinement of $\hat{f}$. The possible refinements depend on the variable being considered:

- For a univariate distribution or a discrete conditional distribution, the algorithm can extend the MoTBF basis with one additional basis function.
- For a continuous conditional density, the algorithm can either perform a partitioning of an existing hyper-cube over the continuous parent variables or include an additional basis function in the MoTBF representation of the local density conditioned on a specific hyper-cube.

A summary specification of the overall algorithm can be found in Algorithm 1. Note that the algorithm relies on an auxiliary function called Refine(M, op) that simply refines the current MoTBF network according to the refinement operation op.

Algorithm 1 The general algorithm for finding an MoTBF representation M of a hybrid Bayesian network B. We use $\mathcal{U}_c$ and $\mathcal{U}_d$ to denote the continuous and the discrete variables in B, respectively.

Input: A hybrid Bayesian network B and an initial MoTBF network M.
Output: An MoTBF representation of the network B.

repeat
  bestGain $\leftarrow -\inf$
  for all $Y \in \mathcal{U}$ **do**
    for all $\boldsymbol{x} \in \Omega_{\text{pa}(Y)\cap\mathcal{U}_d}$ **do**
      (gain, op) $\leftarrow$ EstimateGain$(B, M, Y, \boldsymbol{x})$
      if gain $>$ bestGain **then**
        bestGain $\leftarrow$ gain and bestOp $\leftarrow$ op
      end if
    end for
  end for
  Refine(M, bestOp)
until bestGain $\leq$ threshold
return M

6 Experimental results

In order to validate the proposed translation method, we have performed two empirical studies. The purpose of the first study is to investigate the accuracy that can be obtained using the proposed method. For that we have empirically compared our method with MTE-approximations of standard univariate distributions using the library of translation functions given by Cobb et al. (2006). In the second study we want to demonstrate the possibility of making an online trade-off between accuracy and complexity when translating a Bayesian network into an MoTBF network.

For the first set of experiments, we have used the same set of univariate distributions as in (Cobb et al., 2006) together with the same support sets. Since the purpose of the experiment is to compare the accuracy of the proposed method with the translations by Cobb et al. (2006), we vary the number of MoTBF parameters (i.e., basis functions) between 1 and the maximum number of parameters used by Cobb et al. (2006). From this set we choose the MoTBF approximation that minimizes the KL-divergence between the true distribution

Algorithm 2 EstimateGain$(B, M, Y, \boldsymbol{x})$

Input: A BN B, an MoTBF M representing B, a variable Y whose probability distribution should be refined, and a configuration $\boldsymbol{x}$ over the discrete parents of Y.

Output: A tuple (gain, op), where the first component is the gain of refining M according to the operation specified by op.

Let f be the potential of Y given $\boldsymbol{x}$ in B and let $\hat{f}_Y$ be the representation of f in M.
if $Y \in \mathcal{U}_c$ and $\text{pa}(Y) \cap \mathcal{U}_c \neq \emptyset$ **then**
  Let $\mathcal{P}$ be a partitioning of $\text{pa}(Y) \cap \mathcal{U}_c$
  $\text{gain} \leftarrow -\inf$
  for all $\mathcal{I} \in \mathcal{P}$ **do**
    Let $\hat{f}_s$ be the MoTBF potential obtained from $\hat{f}$ by splitting $\mathcal{I}$ into $\mathcal{I}_0$ and $\mathcal{I}_1$.
    $\text{gain}_s \leftarrow h(\hat{f}_s, f, \hat{f})$
    Let $\hat{f}_a$ be the MoTBF potential obtained from $\hat{f}$ by adding a basis function in $\mathcal{I}$.
    $\text{gain}_a \leftarrow h(\hat{f}_a, f, \hat{f})$
    if $\max_i(\text{gain}_i) \geq \text{gain}$ **then**
      $\text{gain} \leftarrow \max_i(\text{gain}_i)$ and $\text{op} \leftarrow (Y, \arg\max_i(gain_i), \mathcal{I})$
    end if
  end for
else if $Y \in \mathcal{U}_c$ or $(Y \in \mathcal{U}_d$ and $\text{pa}(Y) \cap \mathcal{U}_c \neq \emptyset)$ **then**
  Let $\hat{f}'$ be the MoTBF potential obtained from $\hat{f}$ by adding a basis function.
  $\text{gain} \leftarrow h(\hat{f}', f, \hat{f})$ and $\text{op} \leftarrow (Y, a, \Omega_{\text{pa}(Y)\cap\mathcal{U}_c})$
end if
return (gain, op)

and the MoTBF approximation. In principle, the KL divergence should decrease as the number of parameters increase, but in our experiments this turned out to not necessarily be the case; two possible reasons for this are the numerical instability encountered when using higher-order basis functions as well as the use of numerical methods for evaluating the inner products between the basis functions and the densities to be approximated. The result of the experiment is summarized in Table 1, where we see that the proposed method achieves accuracy results that are either comparable or significantly better than those of Cobb et al. (2006). It should be noted that no split points are introduced in any of the translations, except to define the end-points of the intervals for which the MoTBF distributions have positive support. As a last qualitative remark for this comparison, we note that the proposed method can immediately be applied to other distributions without any modifications to the algorithm, something which is generally not the case when using the method by Cobb et al. (2006). We have also conducted a similar comparison with the MoP-expansion by Shenoy and West (2011), where a Taylor-series expansion around the midpoint of each interval was built. To examine the effectiveness of this translation, we added polynomial terms until the accuracy of the MoTBFs (MoP)-approach was reached. The Taylor expansion required at least as many parameters as the proposed method in all cases, and typically around twice as many parameters needed to be employed.[2]

For the MoTBF (MoP) results in Table 1, the polynomial basis (corresponding to shifted Legendre polynomials) have been calculated recursively and afterwards shifted to the appropriate interval (Kreyzig, 1978). In comparison, the exponential basis has been calculated directly following the Gram-Schmidt process and using numerical integration to evaluate the integrals. Not surprisingly the former approach is less susceptible to numerical instability, which can also account for the differences in accuracy that we observe when comparing the two MoTBF-based methods. Finally, it should be emphasized that the accuracy results obtained for two different distributions are generally not comparable, since the intervals over which the distributions are defined may differ.

For the second experiment we consider a simple parameterized Bayesian network consisting of two nodes $X \sim N(0,1)$ and $Y|\{X = x\} \sim N(w \cdot x, \sigma)$. To investigate how the algorithm proceeds with the translation of a model, we have applied Algorithm 1 (using a polynomial basis) to the model $X \sim N(0,1)$ and $Y|\{X = x\} \sim N(1 \cdot x, 1)$ for 5, 10, 15, and 25 iterations. The results of the experiments can be seen in Figure 7. Observe how the insertion of split points is initially focused on areas with high probability mass whereas areas with lower probability mass are only gradually refined at later iterations.

By varying w and σ we can change the correlation between X and Y. The change in correlation should also affect the sequence in which Algorithm 1 considers the child and parent distributions as well as whether a basis function is added or a split point is being introduced. The effect is illustrated in Figure 8, where we see the result of running the algorithm for 15 iterations for three networks with a varying degree of correlation between the two variables. Notice that as the correlation between X and Y increases, more split points are introduced to encode the correlation (with the result that fewer basis function are being added to model the local behaviors of the distributions). In particular, we see that with $w = 1$ and $\sigma = 0.5$, the

[2]We were only able to perform the test for the Normal, Gamma, and Beta distributions. Matlab's symbolic module could not generate a sufficiently descriptive Taylor-series approximation for the LogNormal distribution within 8hrs CPU-time (Matlab 2011b, running on a 2.8 GHz Intel Core 2 Duo processor with 4GB RAM). The parameter overhead varied from zero (both methods finding the exact representation of Beta(2,2)) to 164 extra parameters required in case of Beta(1.3, 2.7).

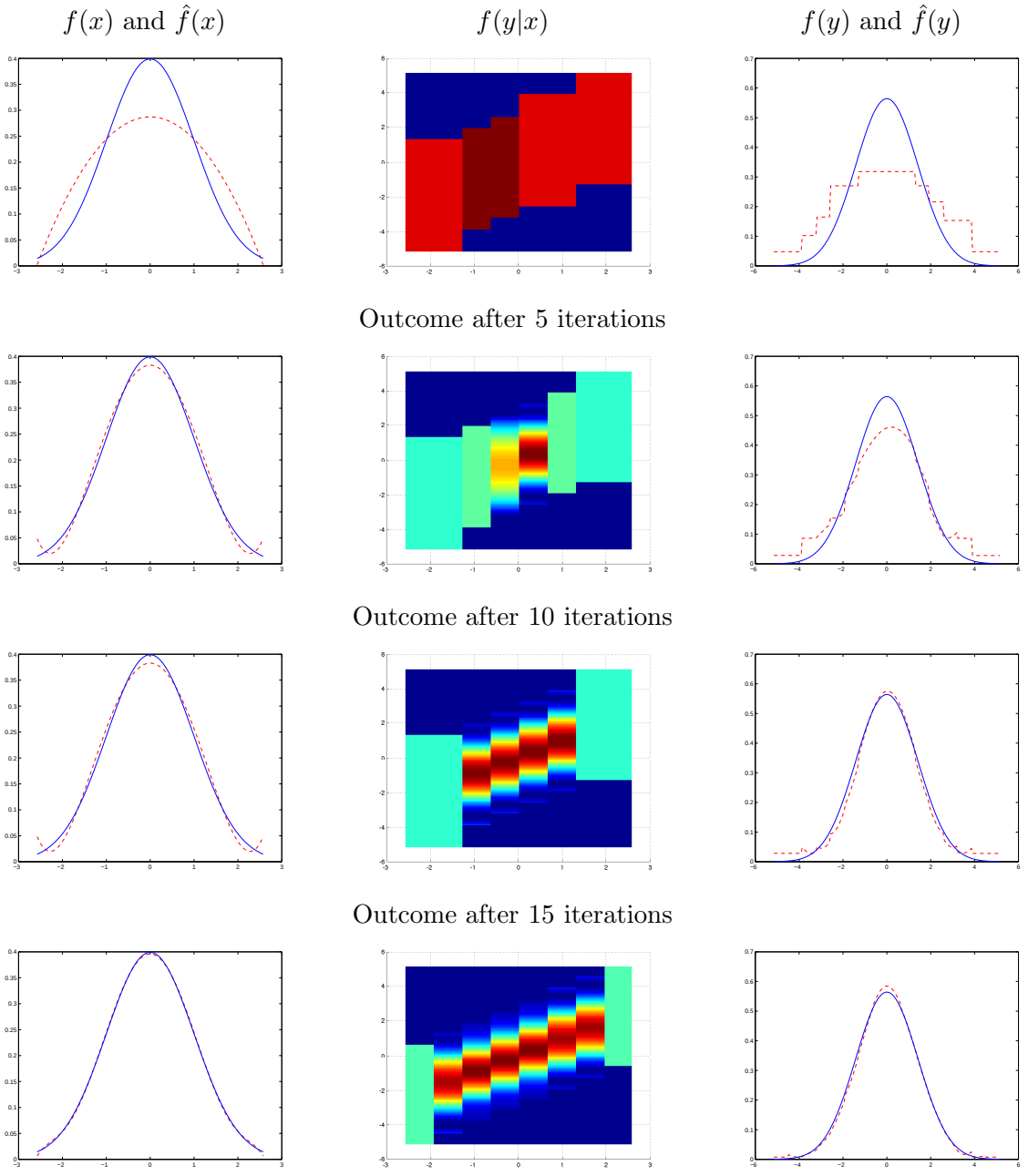


Figure 7: The figure shows the results of applying Algorithm 1 on the model $X \sim N(0,1)$ and $Y \sim N(1 \cdot x, 1)$ for varying number of iterations.

Distribution	Cobb et al.	MoTBFs (MTE)	MoTBFs (MoP)	[Interval]
$N(0,1)$	[15] $3.46 \cdot 10^{-04}$	[8] $1.93 \cdot 10^{-04}$	[14] $2.68 \cdot 10^{-09}$	$[-3,3]$
Gamma(6,1)	[23] $2.10 \cdot 10^{-03}$	[20] $6.41 \cdot 10^{-03}$	[18] $2.91 \cdot 10^{-09}$	$[0.527864, 18.4164]$
Gamma(8,1)	[23] $8.56 \cdot 10^{-04}$	[7] $5.37 \cdot 10^{-03}$	[18] $7.26 \cdot 10^{-11}$	$[1.7085, 22.8745]$
Gamma(11,1)	[23] $2.83 \cdot 10^{-04}$	[12] $6.58 \cdot 10^{-04}$	[17] $1.47 \cdot 10^{-09}$	$[2.09431, 22.6491]$
Beta(2,2)	[17] $2.62 \cdot 10^{-06}$	[9] $3.98 \cdot 10^{-05}$	[3] $1.64 \cdot 10^{-16}$	$[0,1]$
Beta(2.7,1.3)	[17] $3.30 \cdot 10^{-04}$	[8] $5.33 \cdot 10^{-04}$	[9] $4.44 \cdot 10^{-04}$	$[0,1]$
Beta(1.3,2.7)	[17] $3.30 \cdot 10^{-04}$	[8] $5.33 \cdot 10^{-04}$	[16] $4.97 \cdot 10^{-05}$	$[0,1]$
LogNormal(0,0.25)	[23] $3.30 \cdot 10^{-04}$	[19] $9.05 \cdot 10^{-03}$	[18] $3.97 \cdot 10^{-08}$	$[0.22313, 4.4817]$
LogNormal(0,0.5)	[23] $9.90 \cdot 10^{-05}$	[22] $5.69 \cdot 10^{-03}$	[21] $1.86 \cdot 10^{-05}$	$[0.11987, 8.3421]$
LogNormal(0,1)	[23] $6.47 \cdot 10^{-03}$	[6] $4.05 \cdot 10^{-02}$	[17] $3.06 \cdot 10^{-03}$	$[0.0497871, 20.0855]$

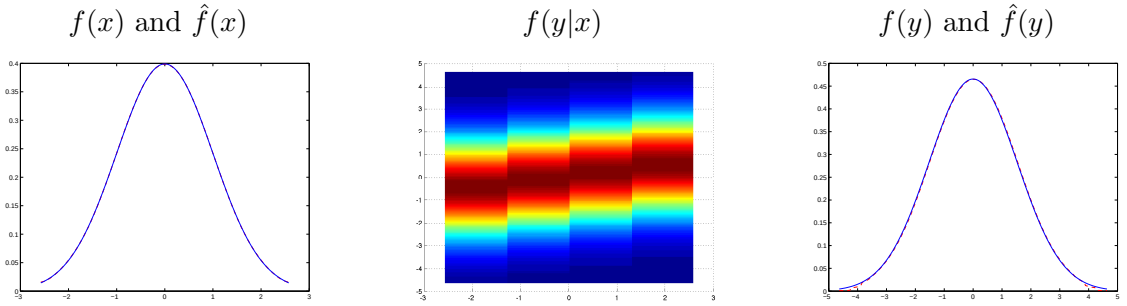
Table 1: The table lists the KL-divergence between the true distribution and the approximation obtained using *i*) the method by Cobb et al. (2006), *ii*) the MoTBF method with exponential basis functions, and *iii*) the MoTBF with polynomial basis functions. The numbers in brackets are the number of parameters used by the approximations.

MoTBF representation of the conditional distribution reduces to a standard discretization.

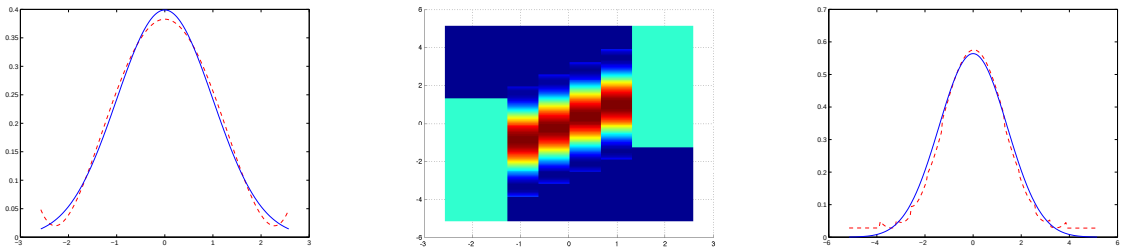
7 Conclusions

In this paper we have proposed a new framework for representing general hybrid Bayesian networks, denoted *mixtures of truncated basis functions*, which supports efficient inference using the Shenoy-Shafer architecture. We have investigated how generalized Fourier series approximation theory and convex optimization techniques can be combined to obtain MoTBF distributions that can approximate univariate probability distributions to any preset quality constraint. We have also discussed how the same methods are viable to handle conditional distribution functions. The translation method is faster and more flexible than existing MTE methods, and it supports an online/anytime trade-off between the accuracy and the complexity of the approximation.

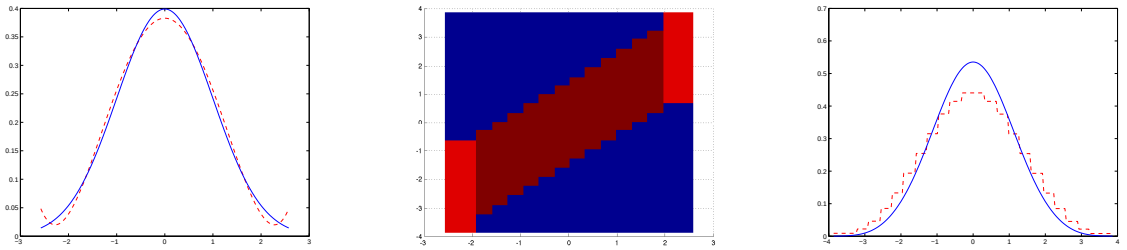
Our future research on this topic will follow three distinct paths: *i*) The translation from the original model to the MoTBF representation is performed off-line, before any evidence is entered into the model. However, a translation that is made to be cost-efficient before entering evidence may not be as effective after evidence has been taken into account. We will therefore consider a *dynamic translation* process in the spirit of Kozlov and Koller (1997) to iteratively re-define an optimal translation given evidence. *ii*) The translation procedure (outlined in Algorithm 1) uses the immediate decrease in KL-divergence per additional parameter introduced in the model as a guide for finding the most cost-efficient translation. In the future we will examine other heuristics that better reflect the trade-off between the cost of *inference* and the obtained precision, where the cost could, e.g., be measured in the size of the junction tree representation. This would require the triangulation of a number of almost identical MoTBF models, which also motivates research into *incremental triangulation* of MoTBF models. *iii*) We want to continue our previous work on learning MTEs/MoTBFs from data (Langseth et al., 2009, 2010), and will investigate how the translation of a Bayesian network structure with conditional probability tables represented by non-parametric density estimates learned from data into an MoTBF compares to models learned directly using a maximum likelihood procedure.



An MoTBF representation of the model $X \sim N(0,1)$ and $Y \sim N(0.3 \cdot x, 1.5)$.



An MoTBF representation of the model $X \sim N(0,1)$ and $Y \sim N(1 \cdot x, 1)$.



An MoTBF representation of the model $X \sim N(0,1)$ and $Y \sim N(1 \cdot x, 0.5)$.

Figure 8: MoTBF representations of three Bayesian networks over $X \sim N(0,1)$ and $Y \sim N(w \cdot x, \sigma)$, defined by $(w = 0.3, \sigma = 1.5)$, $(w = 1, \sigma = 1)$, and $(w = 1, \sigma = 0.5)$. The MoTBF models were obtained by running Algorithm 1 for 15 iterations.

Acknowledgments

This work has been supported by a Senior Grant in the frame of the CALL UCM-EEA-ABEL-02-2009 of the Abel Extraordinary Chair (NILS Project), and by the Spanish Ministry of Science and Innovation, through projects TIN2010-20900-C04-02-03 (entitled *Data mining with PGMs: New algorithms and applications*) and by ERDF (FEDER) funds.

References

References

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

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

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991. ISBN 0-471-06259-6.

Robert G. Cowell, Alexander Philip Dawid, Steffen L. Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Statistics for Engineering and Information Sciences. Springer-Verlag, New York, NY, 1999. ISBN 0-387-98767-3.

Antonio Fernández, Helge Langseth, Thomas D. Nielsen, and Antonio Salmerón. Parameter learning in MTE networks using incomplete data. In Petri Myllymäki, Teemu Roos, and Tommi Jaakkola, editors, *Proceedings of the Fifth European Workshop on Probabilistic Graphical Models*, pages 137–144, 2010.

Nir Friedman and Moises Goldszmidt. Discretizing continuous attributes while learning Bayesian networks. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 157–165, San Francisco, CA, 1996. Morgan Kaufmann Publishers.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

Tomas Hrycej. Gibbs sampling in Bayesian networks (research note). *Artificial Intelligence*, 46:351–363, 1990. ISSN 0004-3702. doi: 10.1016/0004-3702(90)90020-Z.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Laurence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

Alexander V. Kozlov and Daphne Koller. Nonuniform dynamic discretization in hybrid networks. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 314–325, 1997.

Erwin Kreyzig. *Introductory Functional Analysis with Applications*. John Wiley & Sons, New York, NY., 1978. ISBN 0-471-03729-X.

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

Helge Langseth, Thomas D. Nielsen, Rafael Rumí, and Antonio Salmerón. Maximum likelihood learning of conditional MTE distributions. In *Tenth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, volume 5590 of *Lecture Notes in Artificial Intelligence*, pages 240–251. Springer-Verlag, Berlin, Germany, 2009.

Helge Langseth, Thomas D. Nielsen, Rafael Rumí, and Antonio Salmerón. Parameter estimation and model selection for mixtures of truncated exponentials. *International Journal of Approximate Reasoning*, 51:485–498, 2010. doi: http://dx.doi.org/10.1016/j.ijar.2010.01.008.

Steffen L. Lauritzen. Propagation of probabilities, means and variances in mixed graphical association models. *Journal of the American Statistical Association*, 87(420):1098–1108, 1992.

Serafín Moral, Rafael Rumí, and Antonio Salmerón. Mixtures of truncated exponentials in hybrid Bayesian networks. In *Sixth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, volume 2143 of *Lecture Notes in Artificial Intelligence*, pages 145–167. Springer-Verlag, Berlin, Germany, 2001.

Glenn R. Shafer and Prakash P. Shenoy. Probability propagation. *Annals of Mathematics and Artificial Intelligence*, 2:327–352, 1990.

Prakash P. Shenoy and James C. West. Inference in hybrid Bayesian networks using mixtures of polynomials. *International Journal of Approximate Reasoning*, 52:641–657, 2011.

Marshall H. Stone. Application of the theory of boolean rings to general topology. *Transactions of the American Mathematical Society*, 41:375–481, 1937.

Robert A. van Engelen. Approximating Bayesian belief networks by arc removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(8):916–920, 1997.

Lieven Vandenberghe and Stephen Boyd. Semidefinite programming. *SIAM Review*, 38:49–95, 1996.

Karl Weierstrass. Über die analytische Darstellbarkeit sogenannter wilkürlicher Funktionen reeller Argumente. In *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin*, pages 633–639 (Erste Mitteilung); 789–805 (Zweite Mitteilung), 1885.

Joe Whittaker. *Graphical models in applied multivariate statistics*. John Wiley & Sons, Chichester, UK, 1990. ISBN 0-471-91750-8.

Assaf J. Zeevi and Ronny Meir. Density estimation through convex combinations of densities: Approximation and estimation bounds. *Neural Networks*, 10:99–109, 1997.