

# Modeling Nonlinear Deterministic Relationships in Bayesian Networks\*

Barry R. Cobb  
brcobb@ku.edu

Prakash P. Shenoy  
pshenoy@ku.edu

University of Kansas School of Business  
1300 Sunnyside Ave., Summerfield Hall  
Lawrence, KS 66045-7585

April 14, 2005

## Abstract

In a Bayesian network with continuous variables containing a variable(s) that is a conditionally deterministic function of its continuous parents, the joint density function for the variables in the network does not exist. Conditional linear Gaussian distributions can handle such cases when the deterministic function is linear and the continuous variables have a multi-variate normal distribution. In this paper, operations required for performing inference with nonlinear conditionally deterministic variables are developed. We perform inference in networks with nonlinear deterministic variables and non-Gaussian continuous variables by using piecewise linear approximations to nonlinear functions and modeling probability distributions with mixtures of truncated exponentials (MTE) potentials.

**Key Words:** Bayesian networks, MTE potentials, inference, CLG models

---

\*Comments and suggestions for improvement are welcome and will be gratefully appreciated.

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Notation and Definitions</b>	<b>3</b>
2.1	Notation . . . . .	3
2.2	Mixtures of Truncated Exponentials . . . . .	4
2.3	Conditional Mass Functions (CMF) . . . . .	5
2.4	Kullback-Leibler (KL) Divergence . . . . .	5
<b>3</b>	<b>Piecewise Linear Approximations to Nonlinear Functions</b>	<b>6</b>
3.1	Dividing the Domain . . . . .	6
3.2	Algorithm for Splitting Regions . . . . .	6
<b>4</b>	<b>Operations with Linear Deterministic Variables</b>	<b>8</b>
<b>5</b>	<b>Examples</b>	<b>9</b>
5.1	Example One . . . . .	9
5.1.1	Two-point approximation . . . . .	11
5.1.2	Four-point approximation . . . . .	12
5.1.3	Eight-point approximation . . . . .	13
5.1.4	Taylor Series Approximation . . . . .	16
5.2	Example Two . . . . .	18
5.2.1	Piecewise Approximation . . . . .	18
5.2.2	Determining the Distribution of $Y$ . . . . .	19
5.3	Example Three . . . . .	21
5.4	Computing Messages . . . . .	23
5.5	Posterior Marginals . . . . .	23
5.6	Entering Evidence . . . . .	23
<b>6</b>	<b>Summary and Conclusions</b>	<b>26</b>

# 1 Introduction

Bayesian networks model knowledge about propositions in uncertain domains using graphical and numerical representations. At the qualitative level, a Bayesian network is a directed acyclic graph where nodes represent variables and the (missing) edges represent conditional independence relations among the variables. At the numerical level, a Bayesian network consists of a factorization of a joint probability distribution into a set of conditional distributions, one for each variable in the network.

An important class of Bayesian networks with continuous variables are those that have conditionally deterministic variables (a variable that is a deterministic function of its parents). Conditional linear Gaussian (CLG) distributions [Lauritzen and Jensen 2001] can handle such cases when the deterministic function is linear and variables are normally distributed. In models with nonlinear deterministic relationships and non-Gaussian distributions, Monte Carlo methods may be required to obtain an approximate solution. General purpose solution algorithms, e.g., the Shenoy-Shafer architecture, have not been adapted to such models, primarily because that the joint density for the variables in models with deterministic variables does not exist and these methods involve propagation of probability densities.

Approximate inference in Bayesian networks with continuous variables can be performed using mixtures of truncated exponentials (MTE) potentials [Moral *et al.* 2001]. Cobb and Shenoy [2004] define operations which allow the distributions of linear deterministic variables to be determined when the continuous variables are modeled with MTE potentials. This allows MTE potentials to be used for inference in any CLG model, as well as other models that have non-Gaussian and conditionally deterministic variables.

The remainder of this paper is organized as follows. Section 2 introduces notation and definitions used throughout the paper. Section 3 describes a method for approximating a nonlinear function with a piecewise linear function. Section 4 defines operations required for inference in Bayesian networks with conditionally deterministic variables. Section 5 contains examples of determining the distributions of nonlinear conditionally deterministic variables. Section 6 summarizes and states directions for future research.

## 2 Notation and Definitions

This section contains notation and definitions used throughout the paper.

### 2.1 Notation

Random variables will be denoted by capital letters, e.g.,  $A, B, C$ . Sets of variables will be denoted by boldface capital letters, e.g.,  $\mathbf{X}$ . All variables are assumed to take values in

continuous state spaces. If  $\mathbf{X}$  is a set of variables,  $\mathbf{x}$  is a configuration of specific states of those variables. The continuous state space of  $\mathbf{X}$  is denoted by  $\Omega_{\mathbf{X}}$ .

In graphical representations, discrete nodes are represented by single-border ovals, continuous nodes are represented by double-border ovals, and nodes that are deterministic functions of their parents are represented by triple-border ovals.

## 2.2 Mixtures of Truncated Exponentials

A mixture of truncated exponentials (MTE) [Moral *et al.* 2001, Rumí 2003] potential has the following definition.

*MTE potential.* Let  $\mathbf{X} = (X_1, \dots, X_n)$  be an  $n$ -dimensional random variable. A function  $\phi : \Omega_{\mathbf{X}} \mapsto \mathcal{R}^+$  is an MTE potential if one of the next two conditions holds:

1. The potential  $\phi$  can be written as

$$\phi(\mathbf{x}) = a_0 + \sum_{i=1}^m a_i \exp \left\{ \sum_{j=1}^n b_i^{(j)} x_j \right\} \quad (1)$$

for all  $\mathbf{x} \in \Omega_{\mathbf{X}}$ , where  $a_i, i = 0, \dots, m$  and  $b_i^{(j)}, i = 1, \dots, m, j = 1, \dots, n$  are real numbers.

2. The domain of the variables,  $\Omega_{\mathbf{X}}$ , is partitioned into hypercubes  $\{\Omega_{\mathbf{X}_1}, \dots, \Omega_{\mathbf{X}_k}\}$  such that  $\phi$  is defined as

$$\phi(\mathbf{x}) = \phi_i(\mathbf{x}) \quad \text{if } \mathbf{x} \in \Omega_{\mathbf{X}_i}, \quad i = 1, \dots, k, \quad (2)$$

where each  $\phi_i, i = 1, \dots, k$  can be written in the form of equation (1) (i.e. each  $\phi_i$  is an MTE potential on  $\Omega_{\mathbf{X}_i}$ ).

In the definition above,  $k$  is the number of *pieces* and  $m$  is the number of exponential *terms* in each piece of the MTE potential. We will refer to  $\phi_i$  as the  $i$ -th piece of the MTE potential  $\phi$  and  $\Omega_{\mathbf{X}_i}$  as the portion of the domain of  $\mathbf{X}$  approximated by  $\phi_i$ . In this paper, all MTE potentials are equal to zero in unspecified regions.

Moral *et al.* [2002] proposes an iterative algorithm based on least squares approximation to estimate MTE potentials from data. Moral *et al.* [2003] describes a method to approximate conditional MTE potentials using a mixed tree structure. Cobb *et al.* [2003] describes a nonlinear optimization procedure used to fit MTE parameters for approximations to standard PDF's, including the uniform, exponential, gamma, beta, and lognormal distributions.

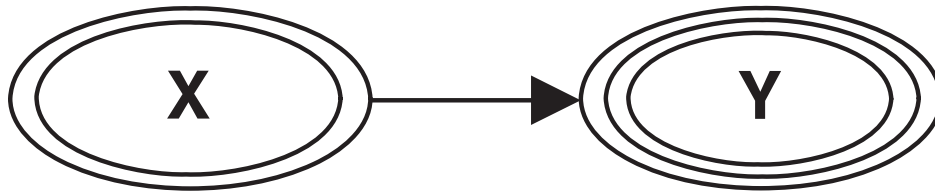


Figure 1: Graphical representation of the conditionally deterministic relationship of  $Y$  given  $X$  determined by the CMF  $p_{Y|x}$ .

### 2.3 Conditional Mass Functions (CMF)

When relationships between continuous variables are deterministic, the joint probability density function (PDF) does not exist. If  $Y$  is a deterministic relationship of variables in  $\mathbf{X}$ , i.e.  $y = g(\mathbf{x})$ , the conditional mass function (CMF) for  $\{Y | \mathbf{x}\}$  is defined as

$$p_{Y|\mathbf{x}} = \mathbf{1}\{y = g(\mathbf{x})\} , \quad (3)$$

where  $\mathbf{1}\{A\}$  is the indicator function of the event  $A$ , i.e.  $\mathbf{1}\{A\}(B) = 1$  if  $B = A$  and 0 otherwise. Graphically, the conditionally deterministic relationship of  $Y$  given  $\mathbf{X}$  is represented in a hybrid Bayesian network model as shown in Figure 1, where  $\mathbf{X}$  consists of a single continuous variable  $X$ .

### 2.4 Kullback-Leibler (KL) Divergence

The relative entropy or Kullback-Leibler (KL) divergence [Kullback and Leibler 1951, MacKay 2003] between  $f_X(x)$  and  $\tilde{f}_X(x)$  is defined as

$$D_{KL}(f_X(x) || \tilde{f}_X(x)) = \int_S f_X(x) \log \frac{f_X(x)}{\tilde{f}_X(x)} dx . \quad (4)$$

Define  $p_{f_{X_i}}$  and  $q_{\tilde{f}_{X_i}}$  as the probability masses of  $f_X(x)$  and  $\tilde{f}_X(x)$ , respectively, in the interval  $(x_{i-1}, x_i]$ . A discrete approximation to the KL divergence statistic over a set of points  $x_i, i = 0, \dots, n$  can be calculated as follows:

$$D'_{KL}(f_X(x) || \tilde{f}_X(x)) = \sum_{i=1}^n p_{f_{X_i}} \log \frac{p_{f_{X_i}}}{q_{\tilde{f}_{X_i}}} . \quad (5)$$

The function  $g(x) = \log(f_X(x)/\tilde{f}_X(x))$  can be interpreted as the information contained in  $x$  for distinguishing between  $f_X(x)$  and  $\tilde{f}_X(x)$ . Thus, KL divergence is the expectation of the information content over the domain  $S$  taken with respect to the distribution  $f_X(x)$ . We use KL divergence to measure the goodness of fit of approximated distributions.

### 3 Piecewise Linear Approximations to Nonlinear Functions

#### 3.1 Dividing the Domain

Suppose that a random variable  $Y$  is a deterministic function of a single variable  $X$ ,  $Y = g(X)$ . The function  $Y = g(X)$  can be approximated by a piecewise linear function. Define a set of ordered points  $x = (x_0, \dots, x_n)$  in the domain of  $X$ , with  $x_0$  and  $x_n$  defined as the endpoints of the domain. A corresponding set of points  $y = (y_0, \dots, y_n)$  is determined by calculating the value of the function  $y = g(x)$  at each point  $x_i$ ,  $i = 0, \dots, n$ . The piecewise linear function (with  $n$  pieces) approximating  $Y = g(X)$  is the function  $Y^{(n)} = g^{(n)}(X)$  defined as follows:

$$g^{(n)}(x) = \begin{cases} \left( y_0 - \frac{y_1 - y_0}{x_1 - x_0} \cdot x_0 \right) + \frac{y_1 - y_0}{x_1 - x_0} \cdot x & \text{if } x_0 \leq x < x_1 \\ \left( y_1 - \frac{y_2 - y_1}{x_2 - x_1} \cdot x_1 \right) + \frac{y_2 - y_1}{x_2 - x_1} \cdot x & \text{if } x_1 \leq x < x_2 \\ \vdots & \vdots \\ \left( y_{n-2} - \frac{y_{n-1} - y_{n-2}}{x_{n-1} - x_{n-2}} \cdot x_{n-2} \right) + \frac{y_{n-1} - y_{n-2}}{x_{n-1} - x_{n-2}} \cdot x & \text{if } x_{n-2} \leq x < x_{n-1} \\ \left( y_{n-1} - \frac{y_n - y_{n-1}}{x_n - x_{n-1}} \cdot x_{n-1} \right) + \frac{y_n - y_{n-1}}{x_n - x_{n-1}} \cdot x & \text{if } x_{n-1} \leq x \leq x_n \end{cases} \quad (6)$$

Let  $g_i^{(n)}(x)$  denote the  $i$ -th piece of the piecewise linear function in (6), e.g.

$$g_1^{(n)}(x) = \left( y_0 - \frac{y_1 - y_0}{x_1 - x_0} \cdot x_0 \right) + \frac{y_1 - y_0}{x_1 - x_0} \cdot x \ .$$

We refer to  $g^{(n)}$  as an  $n$ -point (piecewise linear) approximation of  $g$ . In this paper, all piece-wise linear functions equal zero in unspecified regions.

When the distribution of  $X$  is approximated by an MTE potential, the set of ordered points  $x = (x_0, \dots, x_n)$  will include all endpoints in the domains of each *piece* of the MTE potential (which includes the endpoints of the domain of the variable), all points where concavity of the function  $y = g(x)$  changes, and any extreme points in the domain of the function. Additional points may be included in  $x = (x_0, \dots, x_n)$  to improve the piecewise linear approximation. If a variable is a deterministic function of multiple variables, the definition in (6) can be extended by dividing the domain of the parent variables into hypercubes and creating an approximation of each function in each hypercube.

#### 3.2 Algorithm for Splitting Regions

An initial piecewise approximation is defined (minimally) by splitting the domain of  $X$  at extreme points, points of change in concavity and convexity, and endpoints of pieces of the MTE potential for  $X$ . This initial set of bounds on the pieces of the approximation is defined

as  $x = (x_0^S, \dots, x_\ell^S)$ . The absolute value of the difference between the approximation and the function will increase, then eventually decrease within each region of the approximation. This is due to the fact that the approximation in (6) always lies “inside” the actual function.

Additional pieces may be added to improve the fit between the nonlinear function and the piecewise approximation. Define an allowable error bound,  $\epsilon$ , for the distance between the function  $g(x)$  and its piecewise linear approximation. Define an interval  $\eta$  used to select the next point at which to test the distance between  $g(x)$  and the piecewise approximation. The piecewise linear approximation in (6) is completely defined by the sets of points  $x = (x_0, \dots, x_n)$  and  $y = (y_0, \dots, y_n)$ . The following procedure in pseudo-code determines the sets of points  $x$  and  $y$  which define the piecewise linear approximation when a deterministic variable has one parent.

INPUT :=  $x_0^S, \dots, x_\ell^S, g(x), \epsilon, \eta$

OUTPUT :  $x = (x_0, \dots, x_n), y = (y_0, \dots, y_n)$

INITIALIZATION

$x \leftarrow \{(x_0^S, \dots, x_\ell^S)\}$  /\* Endpoints of MTE pieces, extrema, and inflection points in  $\Omega_{\mathbf{X}}$  \*/

$y \leftarrow \{(g(x_0^S), \dots, g(x_\ell^S))\}$

$i = 0$  /\* Index for the intervals in the domain of  $X$  \*/

DO WHILE  $i < |x|$  /\* Continue until all intervals are refined\*/

$j = 1$  /\* Index for number of test points in an interval \*/

$a = 0$  /\* Previous distance between  $g(x)$  and approximation\*/

$b = 0$  /\* Current distance between  $g(x)$  and approximation \*/

FOR  $j = 1 : (x_{i+1} - x_i) / \eta$

$b = g(x_i + (j - 1) \cdot \eta) - \left( \left( y_i - \frac{y_{i+1} - y_i}{x_{i+1} - x_i} \cdot x_i \right) + \frac{y_{i+1} - y_i}{x_{i+1} - x_i} \cdot (x_i + (j - 1) \cdot \eta) \right)$

IF  $|b| \geq a$  /\* Compare current and previous distance \*/

$a = |b|$  /\*Distance increased; test next point \*/

ELSE

BREAK /\*Distance did not increase; break loop \*/

END IF

END FOR

IF  $a > \epsilon$  /\*Test max. distance versus allowable error bound \*/

$x \leftarrow \text{Rank}(x \cup \{x_i + (j - 2) \cdot \eta\})$  /\* Update  $x$  and re-order ascending \*/

$y \leftarrow \text{Rank}(y \cup \{g(x_i + (j - 2) \cdot \eta)\})$  /\* Update  $y$  and re-order ascending \*/

END IF

$i = i + 1$

END DO

The algorithm refines the piece-wise approximation to the function  $y = g(x)$  until the maximum distance between the function and the piece-wise approximation is no larger than

the specified error bound. A smaller error bound,  $\epsilon$ , produces more pieces in the linear approximation and a closer fit in the theoretical and approximate density functions for the deterministic variable (see, e.g., Section 5.1). A closer approximation using more pieces, however, requires greater computational expense in the inference process.

## 4 Operations with Linear Deterministic Variables

Consider a random variable  $Y$  which is a monotonic function,  $Y = g(X)$ , of a random variable  $X$ . The joint cumulative distribution function (CDF) for  $\{X, Y\}$  is given by  $F_{X,Y}(x, y) = F_X(g^{-1}(y))$  if  $g(X)$  is monotonically increasing and  $F_{X,Y}(x, y) = F_X(x) - F_X(g^{-1}(y))$  if  $g(X)$  is monotonically decreasing. The CDF of  $Y$  is determined as  $F_Y(y) = \lim_{x \rightarrow \infty} F_{X,Y}(x, y)$ . Thus,  $F_Y(y) = F_X(g^{-1}(y))$  if  $g(X)$  is monotonically increasing and  $F_Y(y) = 1 - F_X(g^{-1}(y))$  if  $g(X)$  is monotonically decreasing. By differentiating the CDF of  $Y$ , the PDF of  $Y$  is obtained as

$$f_Y(y) = \frac{d}{dy} F_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} (g^{-1}(y)) \right| , \quad (7)$$

when  $Y = g(X)$  is monotonic. If  $Y$  is a conditionally deterministic linear function of  $X$ , i.e.  $Y = g(x) = ax + b$ ,  $a \neq 0$ , the following operation can be used to determine the marginal PDF for  $Y$ :

$$f_Y(y) = \frac{1}{|a|} \cdot f_X \left( \frac{y - b}{a} \right) . \quad (8)$$

The following definition extends the operation defined in (8) to accommodate piecewise linear functions. Suppose  $Y$  is a conditionally deterministic piecewise linear function of  $X$ ,  $Y = g(X)$ , where  $g_i(x) = a_i x + b_i$ , with each  $a_i \neq 0$ ,  $i = 1, \dots, n$ . Assume the PDF for  $X$  is an MTE potential  $\phi$  with  $k$  pieces, where the  $j$ -th piece is denoted  $\phi_j$  for  $j = 1, \dots, k$ . Let  $n_j$  denote the number of linear segments of  $g$  that intersect with the domain of  $\phi_j$  and notice that  $n = n_1 + \dots + n_j + \dots + n_k$ . The CMF  $p_{Y|x}$  represents the conditionally deterministic relationship of  $Y$  on  $X$ . The following definition will be used to determine the marginal PDF for  $Y$  (denoted  $\chi = (\phi \otimes p_{Y|x})^{\downarrow Y}$ ):

$$\chi(y) = (\phi \otimes p_{Y|x})^{\downarrow Y}(y) \triangleq \begin{cases} 1/a_1 \cdot \phi_1((y - b_1)/a_1) & \text{if } y_0 \leq y < y_1 \\ 1/a_2 \cdot \phi_1((y - b_2)/a_2) & \text{if } y_1 \leq y < y_2 \\ \vdots & \vdots \\ 1/a_{n_1} \cdot \phi_1((y - b_{n_1})/a_{n_1}) & \text{if } y_{n_1-1} \leq y < y_{n_1} \\ \vdots & \vdots \\ 1/a_n \cdot \phi_k((y - b_n)/a_n) & \text{if } y_{n-1} \leq y < y_n \end{cases} , \quad (9)$$

with  $\phi_j$  being the piece of  $\phi$  whose domain is a superset of the domain of  $g_i$ . The normalization constants for each piece of the resulting MTE potential ensures that the CDF of the

resulting MTE potential matches the CDF of the theoretical MTE potential at the endpoints of the domain of the resulting PDF. From Theorem 3 in [Cobb and Shenoy 2004], it follows that the class of MTE potentials is closed under the operation in (9); thus, the operation can be used for inference in Bayesian networks with deterministic variables<sup>1</sup>. Note that the class of MTE potentials is not closed under the operation in (7), which is why we approximate nonlinear functions with piece-wise linear functions.

## 5 Examples

The following examples illustrate determination of the distributions of random variables which are nonlinear deterministic functions of their parents, as well as inference in a simple Bayesian network with a nonlinear deterministic variable.

### 5.1 Example One

Suppose  $X$  is normally distributed with a mean of 4 and a standard deviation of 1, i.e.  $X \sim N(4, 1^2)$ , and  $Y$  is a conditionally deterministic function of  $X$ ,  $y = g(x) = x^2$ . The distribution of  $X$  is modeled with an two-piece, three-term MTE potential as described by Cobb and Shenoy [2003]. Since the MTE approximation to the normal distribution has domain  $[\mu - 3\sigma, \mu + 3\sigma]$ ,  $X$  has domain  $[1, 7]$  and is defined as follows:

$$\phi(x) = P(X) = \begin{cases} -0.010593 + 197.589211 \exp\{2.2568434(x - 4)\} \\ -462.688510 \exp\{2.3434117(x - 4)\} + 265.509914 \exp\{2.4043270(x - 4)\} \\ \quad \text{if } 1 \leq x < 4 \\ -0.010593 + 197.589211 \exp\{-2.2568434(x - 4)\} \\ -462.688510 \exp\{-2.3434117(x - 4)\} + 265.509914 \exp\{-2.4043270(x - 4)\} \\ \quad \text{if } 4 \leq x \leq 7 \end{cases}$$

Recall from Section 2.2, we can denote, e.g.,

$$\phi_1(x) = -0.010593 + 197.589211 \exp\{2.2568434(x - 4)\} \\ -462.688510 \exp\{2.3434117(x - 4)\} + 265.509914 \exp\{2.4043270(x - 4)\} ,$$

with  $\Omega_{X_1} = \{x : 1 \leq x < 4\}$

A graphical representation of two-point, four-point, and eight-point linear approximations to the function  $y = g(x) = x^2$  are shown in Figures 2, 3, and 4.

---

<sup>1</sup>Marginalization of some MTE potentials requires replacement of a linear term with an MTE approximation so that the result is an MTE potential. For details, see [Cobb and Shenoy 2003].

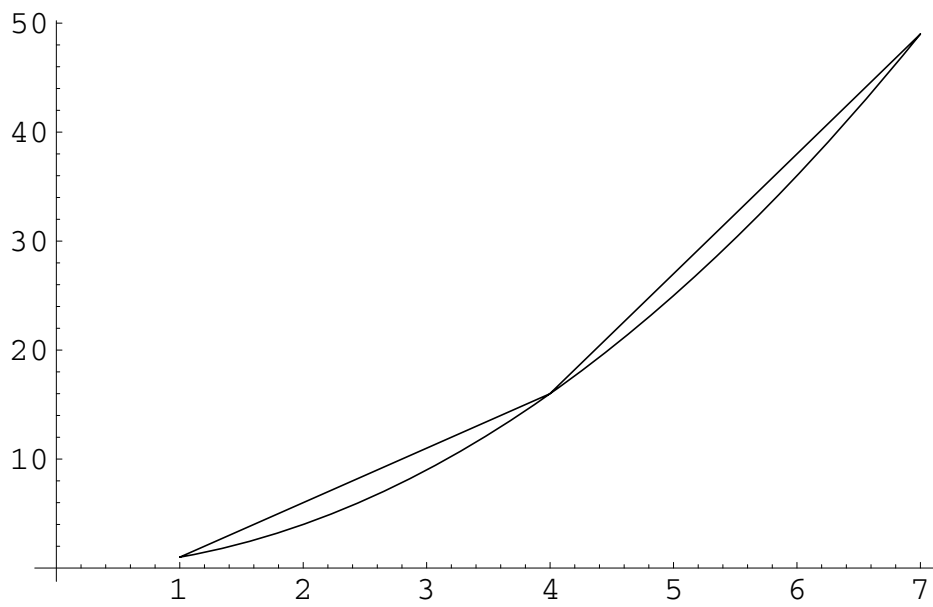


Figure 2: Graphical representation of the two-point linear approximation to the function  $y = g(x) = x^2$ , overlaid on the actual function.

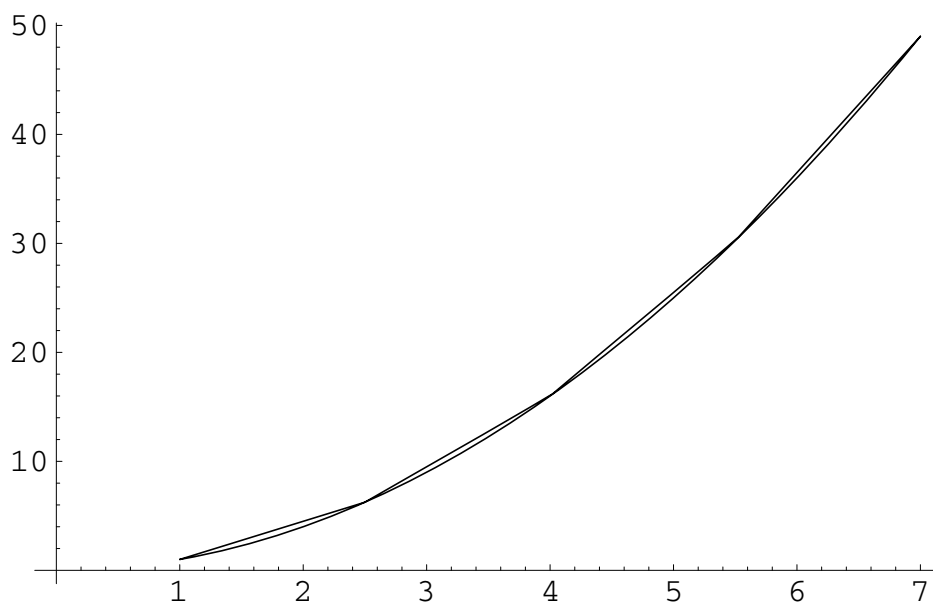


Figure 3: Graphical representation of the four-point linear approximation to the function  $y = g(x) = x^2$ , overlaid on the actual function.

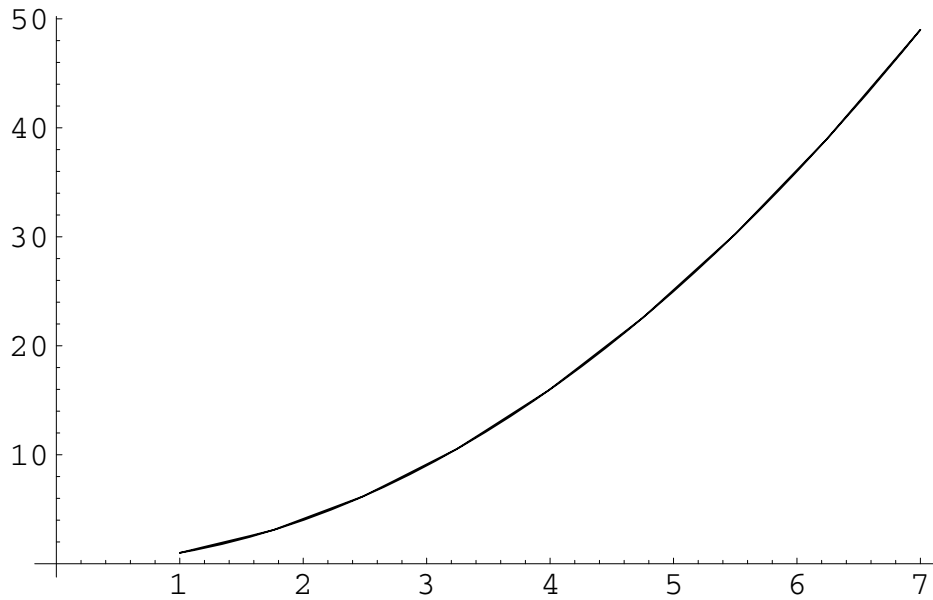


Figure 4: Graphical representation of the eight-point linear approximation to the function  $y = g(x) = x^2$ , overlaid on the actual function.

### 5.1.1 Two-point approximation

The function  $y = g(x) = x^2$  does not have any changes in concavity over the interval  $[1, 7]$ , so the two-point linear approximation has  $x_0 = 1$ ,  $x_1 = 4$ ,  $x_2 = 7$ ,  $y_0 = 1$ ,  $y_1 = 16$ , and  $y_2 = 49$ , so the function representing the two-piece linear approximation is defined as

$$g^{(2)}(x) = \begin{cases} 5x - 4 & \text{if } 1 \leq x < 4 \\ 11x - 28 & \text{if } 4 \leq x \leq 7 \end{cases} \quad (10)$$

The conditional distribution for  $Y$  is represented by a CMF as follows:

$$\psi^{(2)}(x, y) = p_{Y|x}(y) = \mathbf{1}\{y = g^{(2)}(x)\} .$$

The marginal distribution for  $Y$  is determined by calculating  $\chi^{(2)} = (\phi \otimes \psi^{(2)}) \downarrow^Y$ . The MTE potential for  $Y$  is

$$\chi^{(2)}(y) = \begin{cases} (1/5) \cdot \phi_1(0.2y + 0.8) & \text{if } 1 \leq y < 16 \\ (1/11) \cdot \phi_2(0.0909091y + 2.54545) & \text{if } 16 \leq y \leq 49 . \end{cases}$$

The two-piece MTE approximation to the distribution for  $Y$  is shown graphically in Figure 5, overlaid on the distribution created by using the following transformation [Larsen and Marx 2001]:

$$f_Y(y) = \frac{1}{2\sqrt{y}} (\phi(\sqrt{y}) + \phi(-\sqrt{y})) . \quad (11)$$

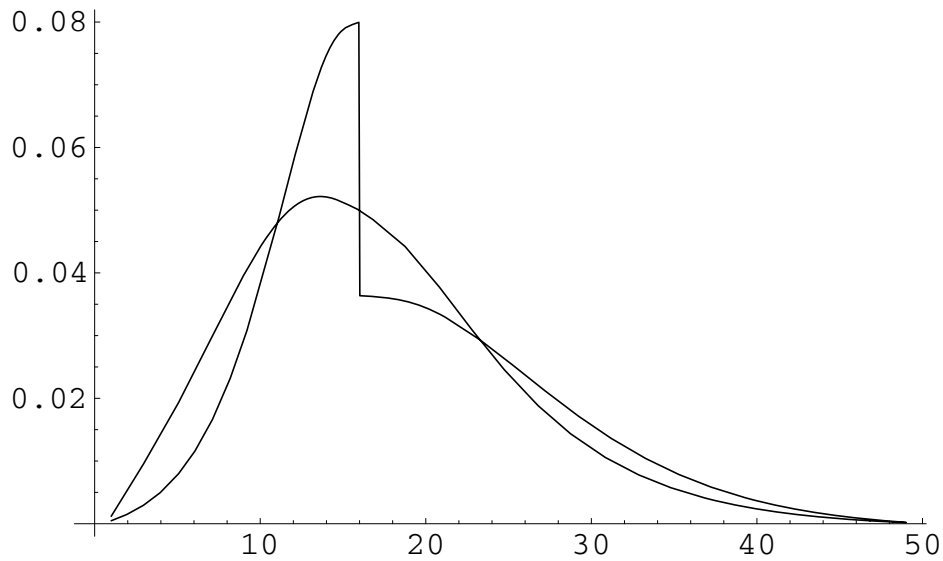


Figure 5: Two-piece MTE approximation to the distribution for  $Y$  overlaid on  $f_Y(y)$ .

Note that while this transformation results in the correct density function for  $Y$ , it does not result in an MTE potential. The CDF's associated with the two-piece MTE approximation and the PDF in (11) are shown in Figure 6.

The KL divergence between  $\chi^{(2)}$  and  $f_Y$  is 0.062418. The MTE approximation satisfies the following probability mass constraint:

$$\int_{x_0}^{x_1} \phi_1(x) dx = \int_{y_0}^{y_1} \chi^{(2)} \chi_2(y) dy = \int_{x_1}^{x_2} \phi_2(x) dx = \int_{y_1}^{y_2} \chi^{(2)}(y) dy = 0.5000 .$$

### 5.1.2 Four-point approximation

The four-point linear approximation is characterized by points  $x = (x_0, \dots, x_4) = (1, 2.5, 4, 5.5, 7)$  and  $y = (y_0, \dots, y_4) = (1, 6.25, 16, 30.25, 49)$ , so the function representing the four-point linear approximation is defined as

$$g^{(4)}(x) = \begin{cases} 3.5x - 2.5 & \text{if } 1 \leq x < 2.5 \\ 6.5x - 10 & \text{if } 2.5 \leq x < 4 \\ 9.5x - 22 & \text{if } 4 \leq x < 5.5 \\ 12.5x - 38.5 & \text{if } 5.5 \leq x \leq 7 . \end{cases}$$

The conditional distribution for  $Y$  is represented by a CMF as follows:

$$\psi^{(4)}(x, y) = p_{Y|x}(y) = \mathbf{1}\{y = g^{(4)}(x)\} .$$

The marginal distribution for  $Y$  is determined by calculating  $\chi^{(4)} = (\phi \otimes \psi^{(4)})^{\downarrow Y}$ . The MTE potential for  $Y$  is

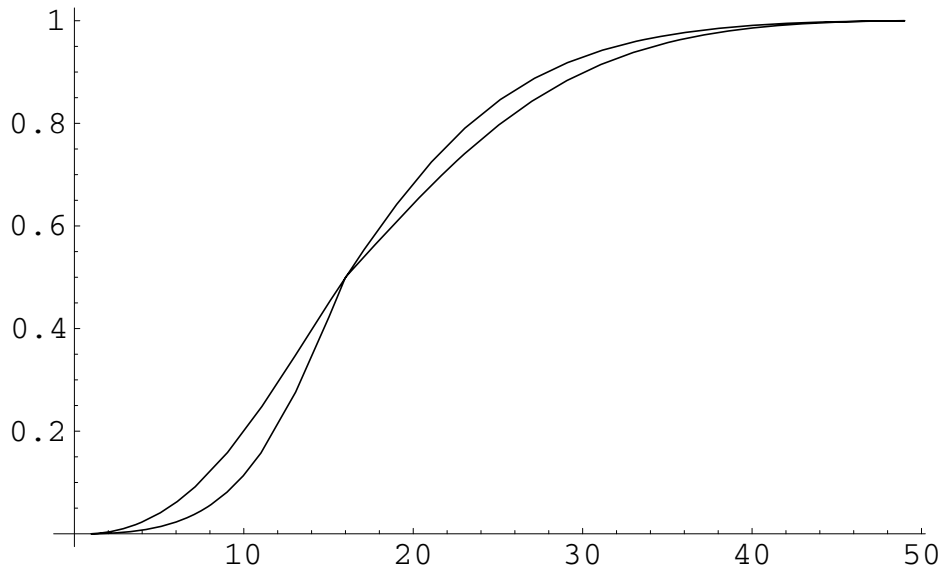


Figure 6: CDF for the two-piece MTE approximation to the distribution for  $Y$  overlaid on the CDF associated with  $f_Y(y)$ .

$$\chi^{(4)}(y) = \begin{cases} (1/3.5) \cdot \phi_1(0.285714y + 0.714286) & \text{if } 1 \leq y < 6.25 \\ (1/6.5) \cdot \phi_1(0.153846y + 1.53846) & \text{if } 6.25 \leq y < 16 \\ (1/9.5) \cdot \phi_2(0.105263y + 2.31579) & \text{if } 16 \leq y < 30.25 \\ (1/12.5) \cdot \phi_2(0.08y + 3.08) & \text{if } 30.25 \leq y \leq 49 . \end{cases}$$

The four-piece MTE approximation to the distribution for  $Y$  is shown graphically in Figure 7, overlaid on the distribution created by using the transformation in (11). The CDF's associated with the four-piece MTE approximation and the PDF in (11) are shown in Figure 8.

The KL divergence between  $\chi^{(4)}$  and  $f_Y$  is 0.009705. The MTE approximation satisfies the following probability mass constraints:

$$\begin{aligned} \int_{x_0}^{x_1} \phi_1(x) dx &= \int_{y_0}^{y_1} \chi^{(4)}(y) dy = \int_{x_3}^{x_4} \phi_2(x) dx = \int_{y_3}^{y_4} \chi^{(4)}(y) dy = 0.0673 , \\ \int_{x_1}^{x_2} \phi_1(x) dx &= \int_{y_1}^{y_2} \chi^{(4)}(y) dy = \int_{x_2}^{x_3} \phi_2(x) dx = \int_{y_2}^{y_3} \chi^{(4)}(y) dy = 0.4327 . \end{aligned}$$

### 5.1.3 Eight-point approximation

The eight-point linear approximation is characterized by points  $x = (x_0, \dots, x_8) = (1, 1.75, 2.5, 3.25, 4, 4.75, 5.5, 6.25, 7)$  and  $y = (y_0, \dots, y_8) = (1, 3.0625, 6.25, 10.5625, 16, 22.5625, 30.25, 39.0625, 49)$ , so the function representing the eight-point linear approximation is defined as

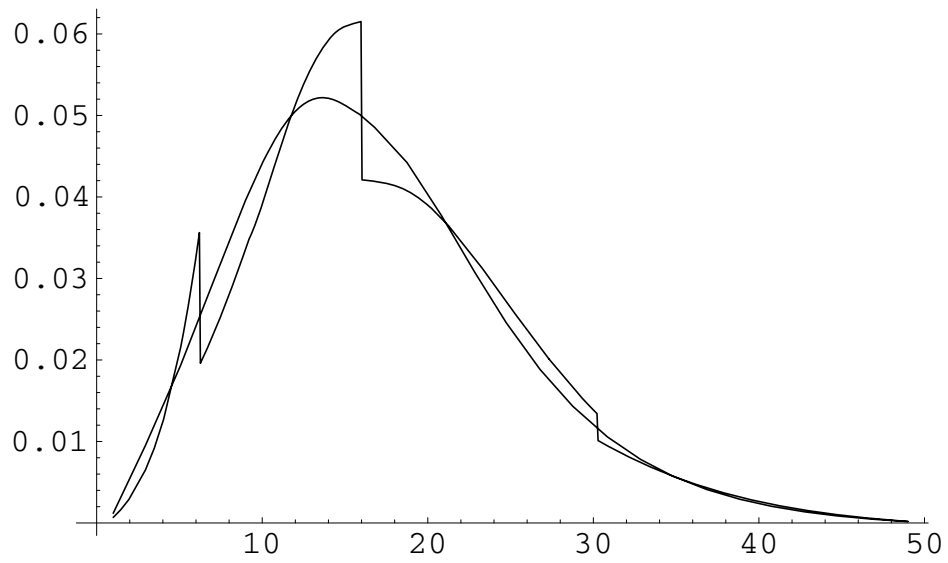


Figure 7: Four-piece MTE approximation to the distribution for  $Y$  overlaid on  $f_Y(y)$ .

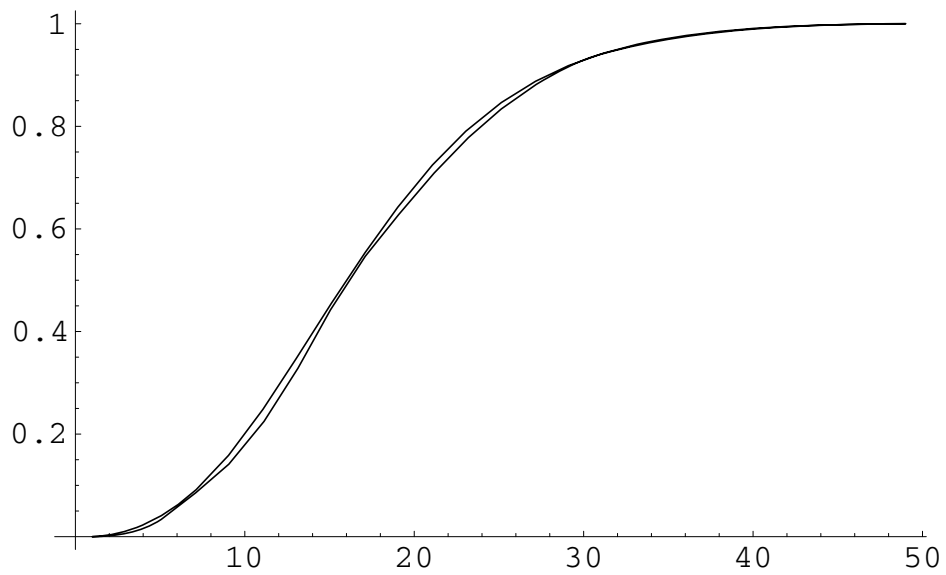


Figure 8: CDF for the four-piece MTE approximation to the distribution for  $Y$  overlaid on the CDF associated with  $f_Y(y)$ .

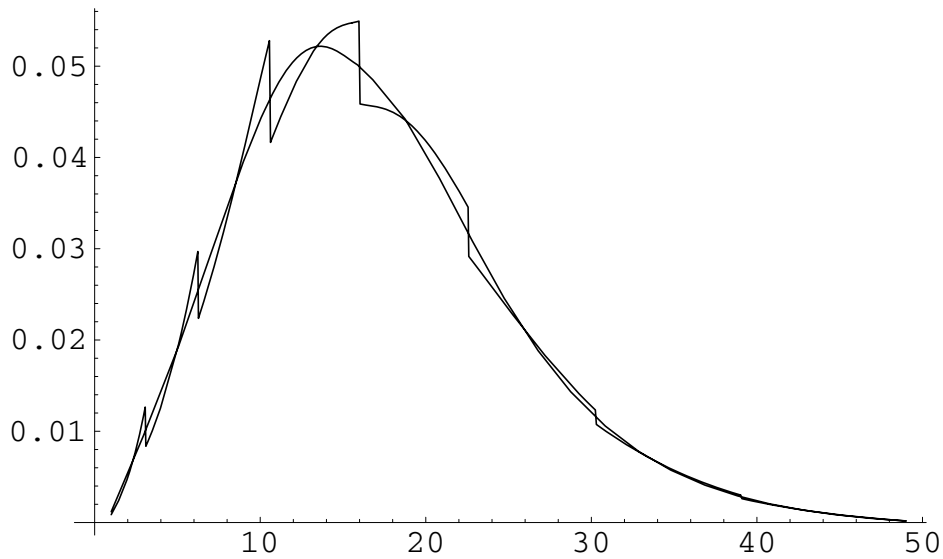


Figure 9: Eight-piece MTE approximation to the distribution for  $Y$  overlaid on  $f_Y(y)$ .

$$g^{(8)}(x) = \begin{cases} 2.75x - 1.75 & \text{if } 1 \leq x < 1.75 \\ 4.25x - 4.375 & \text{if } 1.75 \leq x < 2.5 \\ 5.75x - 8.125 & \text{if } 2.5 \leq x < 3.25 \\ 7.25x - 13 & \text{if } 3.25 \leq x < 4 \\ 8.75x - 19 & \text{if } 4 \leq x < 4.75 \\ 10.25x - 26.125 & \text{if } 4.75 \leq x < 5.5 \\ 11.75x - 34.375 & \text{if } 5.5 \leq x < 6.25 \\ 13.25x - 43.75 & \text{if } 6.25 \leq x \leq 7. \end{cases}$$

The conditional distribution for  $Y$  is represented by a CMF as follows:

$$\psi^{(8)}(x, y) = p_{Y|x}(y) = \mathbf{1}\{y = g^{(8)}(x)\} .$$

The marginal distribution for  $Y$  is determined by calculating  $\chi^{(8)} = (\phi \otimes \psi^{(8)})^{\downarrow Y}$ . The eight-piece MTE approximation to the distribution for  $Y$  is shown graphically in Figure 9, overlaid on the distribution created by using the transformation in (11). The CDF's associated with the eight-piece MTE approximation and the PDF in (11) are shown in Figure 10.

The KL divergence between  $\chi^{(8)}$  and  $f_Y$  is 0.002569. The MTE approximation satisfies the following probability mass constraints for each  $i, j$ ,  $i = 0, \dots, n$ ,  $j = 1, \dots, n$ :

$$\int_{x_i}^{x_j} \phi(x) dx = \int_{y_i}^{y_j} \chi^{(8)}(y) dy .$$

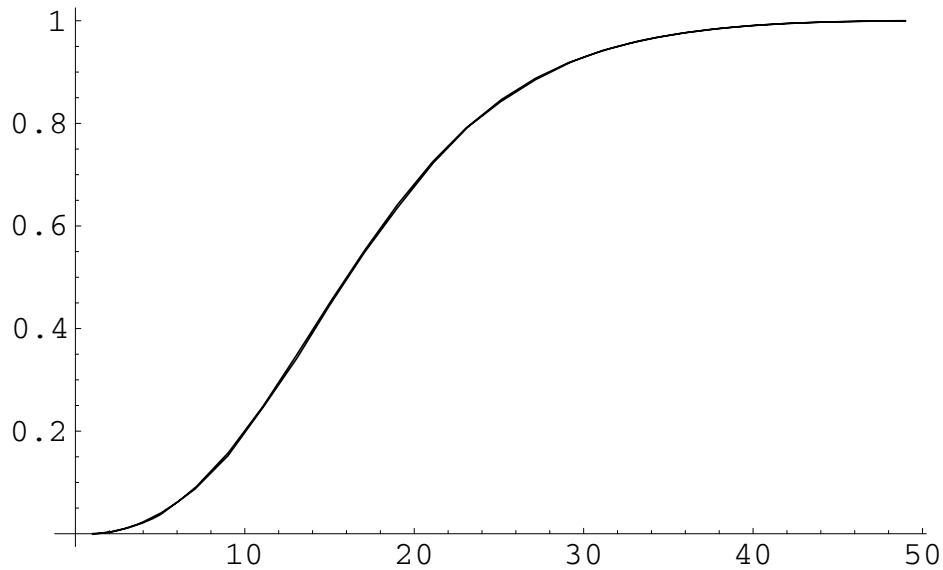


Figure 10: CDF for the eight-piece MTE approximation to the distribution for  $Y$  overlaid on the CDF associated with  $f_Y(y)$ .

#### 5.1.4 Taylor Series Approximation

The two-point, four-point, and eight-point linear approximations to  $y = g(x) = x^2$  used in Example One allow the domain of  $g(x)$  to be the same as its linear approximations  $g^{(2)}(x)$ ,  $g^{(4)}(x)$ , and  $g^{(8)}(x)$ . The approximation is always greater than or equal to a convex function and less than or equal to a concave function. A Taylor series approximation could also be used to approximate the non-linear function in this example. The first-order Taylor series generated by a function  $g(x)$  at  $x = a$  is  $g(a) + g'(a)(x - a)$ . A four-point linear approximation to  $y = g(x) = x^2$  over the interval  $[1, 7]$  is

$$g^T(x) = \begin{cases} 3.5x - 3.0625 & \text{if } 1 \leq x < 2.5 \\ 6.5x - 10.5625 & \text{if } 2.5 \leq x < 4 \\ 9.5x - 22.5625 & \text{if } 4 \leq x < 5.5 \\ 12.5x - 39.0625 & \text{if } 5.5 \leq x \leq 7 \end{cases} .$$

This piecewise approximation is shown in Figure 11, overlaid on  $g(x) = x^2$ .

The Taylor series approximation is always less than or equal to a convex function. In this case,  $g^T(1) = 0.4375$  and  $g^T(7) = 48.4375$ , which means the distribution for  $Y$  resulting from using the piecewise linear approximation will not have the same domain as the function  $g(x)$ . An MTE approximation to the distribution for  $Y$  using the four-point Taylor series approximation is shown in Figure 12. The KL divergence between this distribution and the one generated using the transformation in (11) is 0.007160.

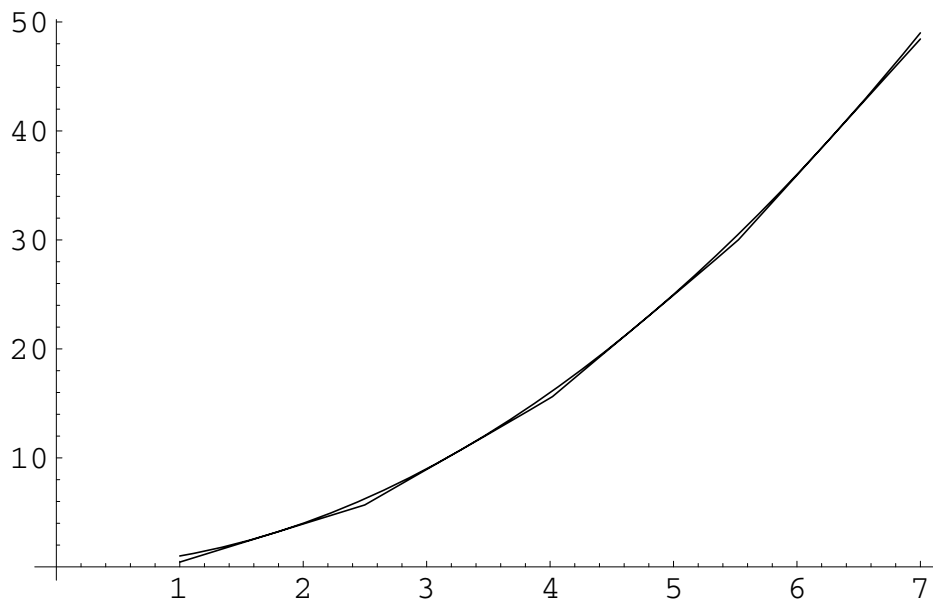


Figure 11: Four-point linear Taylor series approximation to  $g(x) = x^2$ .

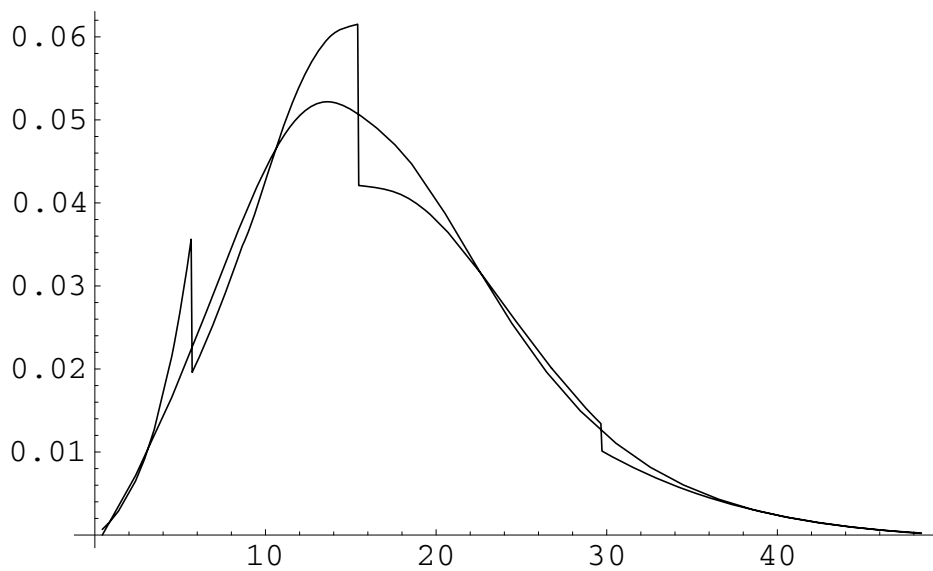


Figure 12: Four-piece MTE approximation to the distribution for  $Y$  overlayed on  $f_Y(y)$ , generated using the Taylor series piecewise approximation.

## 5.2 Example Two

Suppose  $X$  is normally distributed with a mean of 0 and a standard deviation of 1, i.e.  $X \sim N(0, 1^2)$ , and  $Y$  is a conditionally deterministic function of  $X$ ,  $y = g(x) = x^3$ . The distribution of  $X$  is modeled with an two-piece, three-term MTE potential as described in Cobb and Shenoy [2003]. Since the MTE approximation to the normal distribution has domain  $[\mu - 3\sigma, \mu + 3\sigma]$ ,  $X$  has domain  $[-3, 3]$  and is defined as follows:

$$\phi(x) = P(X) = \begin{cases} -0.010593 + 197.589211 \exp\{2.2568434x\} \\ -462.688510 \exp\{2.3434117x\} + 265.509914 \exp\{2.4043270x\} \\ \quad \text{if } -3 \leq x < 0 \\ -0.010593 + 197.589211 \exp\{-2.2568434x\} \\ -462.688510 \exp\{-2.3434117x\} + 265.509914 \exp\{-2.4043270x\} \\ \quad \text{if } 0 \leq x \leq 3 \end{cases}$$

### 5.2.1 Piecewise Approximation

Over the region  $[-3, 3]$ , the function  $y = g(x) = x^3$  has an inflection point at  $x = 0$ , which is also an endpoint of a piece of the MTE approximation to the PDF of  $X$ . To initialize the algorithm in Section 3.2, we define  $x = (x_0^S, x_1^S, x_2^S) = (-3, 0, 3)$  and  $y = (y_0^S, y_1^S, y_2^S) = (-27, 0, 27)$ . For this example, define  $\epsilon = 1$  and  $\eta = 0.06$  (which divides the domain of  $X$  into 100 equal intervals).

The procedure in Section 3.2 terminates after finding sets of points  $x = (x_0, \dots, x_8)$  and  $y = (y_0, \dots, y_8)$  as follows:

$$\begin{aligned} x &= (-3.00, -2.40, -1.74, -1.02, 0.00, 1.02, 1.74, 2.40, 3.00) , \\ y &= (-27.000, -13.824, -5.268, -1.061, 0.000, 1.061, 5.268, 13.824, 27.000) . \end{aligned}$$

The function representing the eight-point linear approximation is defined as

$$g^{(8)}(x) = \begin{cases} g_1^{(8)}(x) & \text{if } x_0 \leq x < x_1 \\ g_2^{(8)}(x) & \text{if } x_1 \leq x < x_2 \\ g_3^{(8)}(x) & \text{if } x_2 \leq x < x_3 \\ g_4^{(8)}(x) & \text{if } x_3 \leq x < x_4 \\ g_5^{(8)}(x) & \text{if } x_4 \leq x < x_5 \\ g_6^{(8)}(x) & \text{if } x_5 \leq x < x_6 \\ g_7^{(8)}(x) & \text{if } x_6 \leq x < x_7 \\ g_8^{(8)}(x) & \text{if } x_7 \leq x \leq x_8 . \end{cases} = \begin{cases} 21.960x + 38.880 & \text{if } -3.00 \leq x < -2.40 \\ 12.964x + 17.289 & \text{if } -2.40 \leq x < -1.74 \\ 5.843x + 4.898 & \text{if } -1.74 \leq x < -1.02 \\ 1.040x & \text{if } -1.02 \leq x < 0 \\ 1.040x & \text{if } 0 \leq x < 1.02 \\ 5.843x - 4.898 & \text{if } 1.02 \leq x < 1.74 \\ 12.964x - 17.289 & \text{if } 1.74 \leq x < 2.40 \\ 21.960x - 38.880 & \text{if } 2.40 \leq x \leq 3.00 . \end{cases}$$

The piecewise linear approximation  $g^{(8)}(x)$  is shown in Figure 13, overlaid on the function  $y = g(x)$ .

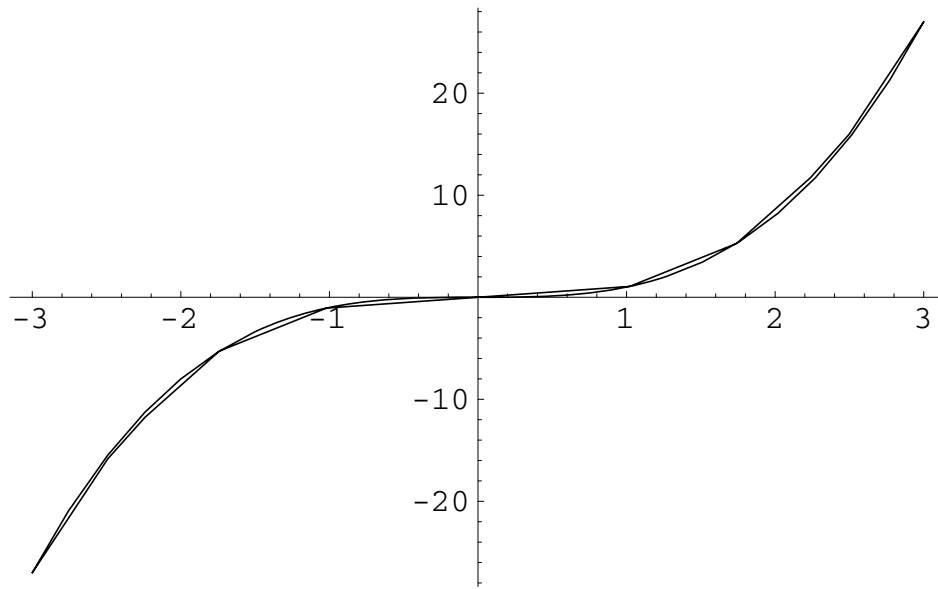


Figure 13: The piecewise linear approximation  $g^{(8)}(x)$  overlaid on the function  $y = g(x)$ .

The conditional distribution for  $Y$  is represented by a CMF as follows:

$$\psi^{(8)}(x, y) = p_{Y|x}(y) = \mathbf{1}\{y = g^{(8)}(x)\} .$$

### 5.2.2 Determining the Distribution of $Y$

The marginal distribution for  $Y$  is determined by calculating  $\chi^{(8)} = (\phi \otimes \psi^{(8)})^{\downarrow Y}$ . The MTE potential for  $Y$  is

$$\chi^{(8)}(y) = \begin{cases} (1/21.960) \cdot \phi^{(1)}(0.0455y - 1.7705) & \text{if } -27.000 \leq y < -13.824 \\ (1/12.964) \cdot \phi_1(0.0771y - 1.3336) & \text{if } -13.824 \leq y < -5.268 \\ (1/5.843) \cdot \phi_1(0.1712y - 0.8384) & \text{if } -5.268 \leq y < -1.061 \\ (1/1.040) \cdot \phi_1(0.9612y) & \text{if } -1.061 \leq y \leq 0.000 \\ (1/1.040) \cdot \phi_2(0.9612y) & \text{if } 0.000 \leq y < 1.061 \\ (1/5.843) \cdot \phi_2(0.1712y + 0.8384) & \text{if } 1.061 \leq y < 5.628 \\ (1/12.964) \cdot \phi_2(0.0771y + 1.3336) & \text{if } 5.628 \leq y < 13.824 \\ (1/21.960) \cdot \phi_2(0.0455y + 1.7705) & \text{if } 13.824 \leq y \leq 27.000 . \end{cases}$$

The MTE potential associated with  $Y$  is shown graphically in Figure 14, overlaid on the distribution created by using the transformation

$$f_Y(y) = f_X(g_1^{-1}(y)) \frac{d}{dy} (g_1^{-1}(y)) . \quad (12)$$

The CDF associated with the eight-piece MTE approximation is shown in Figure 15, overlaid on the CDF associated with the PDF from the transformation in (12).

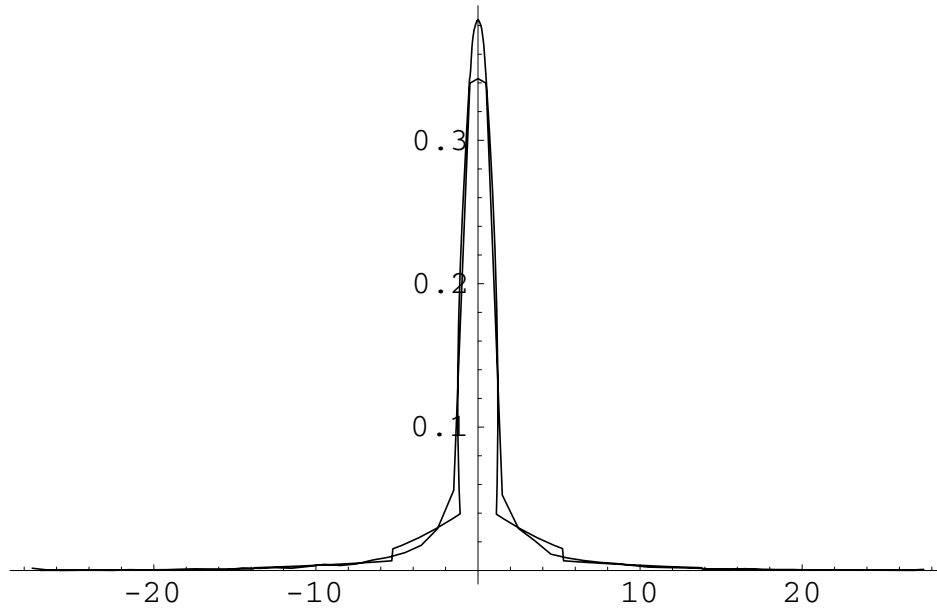


Figure 14: Eight-piece MTE approximation to the distribution for  $Y$  overlaid on the PDF created using the transformation in (12).

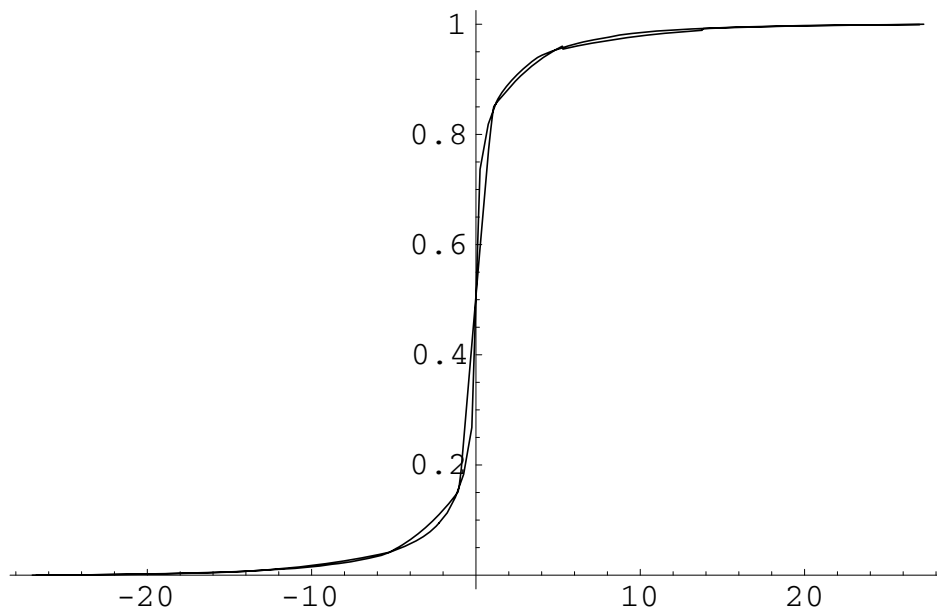


Figure 15: CDF for the eight-piece MTE approximation to the distribution for  $Y$  overlaid on the CDF created using the transformation in (12).

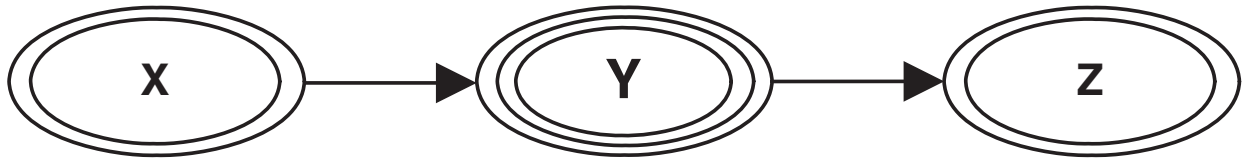


Figure 16: The Bayesian network for Example Three.

### 5.3 Example Three

The Bayesian network in this example (see Figure 16) contains one variable ( $X$ ) with a non-Gaussian potential, one variable ( $Z$ ) with a Gaussian potential, and one variable ( $Y$ ) which is a deterministic linear function of its parent. The probability distribution for  $X$  is a beta distribution, i.e.  $\mathcal{L}(X) \sim \text{Beta}(\alpha = 2.7, \beta = 1.3)$ . The PDF for  $X$  is approximated (using the methods described in [Cobb *et al.* 2003]) by an MTE potential as follows:

$$\phi(x) = P(X) = \begin{cases} -5.951669 + 5.573316 \exp\{0.461388x\} - 0.378353 \exp\{-6.459391x\} & \text{if } 0 < x < d^- \\ 0.473654 - 6.358483 \exp\{-2.639474x\} + 2.729395 \exp\{-0.331472x\} & \text{if } d^- \leq x < m \\ 1.823067 - (5.26E - 12) \exp\{26.000041x\} + 0.035775 \exp\{0.529991x\} & \text{if } m \leq x < 1 \end{cases}$$

where  $m = (1 - \alpha)/(2 - \alpha - \beta) = 0.85$  and

$$d^- = \frac{(\alpha-1)(\alpha+\beta-3) - \sqrt{(\beta-1)(\alpha-1)(\alpha+\beta-3)}}{(\alpha+\beta-3)(\alpha+\beta-2)} = 0.493.$$

The MTE potential for  $X$  is shown graphically in Figure 17, overlaid on the actual  $\text{Beta}(2.7, 1.3)$  distribution.

The variable  $Y$  is a conditionally deterministic function of  $X$ ,  $y = g(x) = -0.5x^3 + x^2$ . The five-point linear approximation is characterized by points  $x = (x_0, \dots, x_5) = (0, 0.220, 0.493, 0.667, 0.850, 1)$  and  $y = (y_0, \dots, y_5) = (0, 0.043, 0.183, 0.296, 0.415, 0.500)$ . The points  $x_0, x_2, x_3$ , and  $x_5$  are defined according to the endpoints of the pieces of  $\phi$ . The point  $x_4$  is an inflection point in the function  $g(x)$  and the point  $x_1 = 0.220$  is found by the algorithm in Section 3.2 with  $\epsilon = 0.015$  and  $\eta = 0.01$ . The function representing the five-piece linear approximation (shown graphically in Figure 18 overlaid on  $g(x)$ ) is defined as

$$g^{(5)}(x) = \begin{cases} 0.1958x & \text{if } 0 \leq x < 0.220 \\ 0.5130x - 0.0698 & \text{if } 0.220 \leq x < 0.493 \\ 0.6516x - 0.1381 & \text{if } 0.493 \leq x < 0.667 \\ 0.6499x - 0.1369 & \text{if } 0.667 \leq x < 0.850 \\ 0.5638x - 0.0638 & \text{if } 0.850 \leq x \leq 1 \end{cases}$$

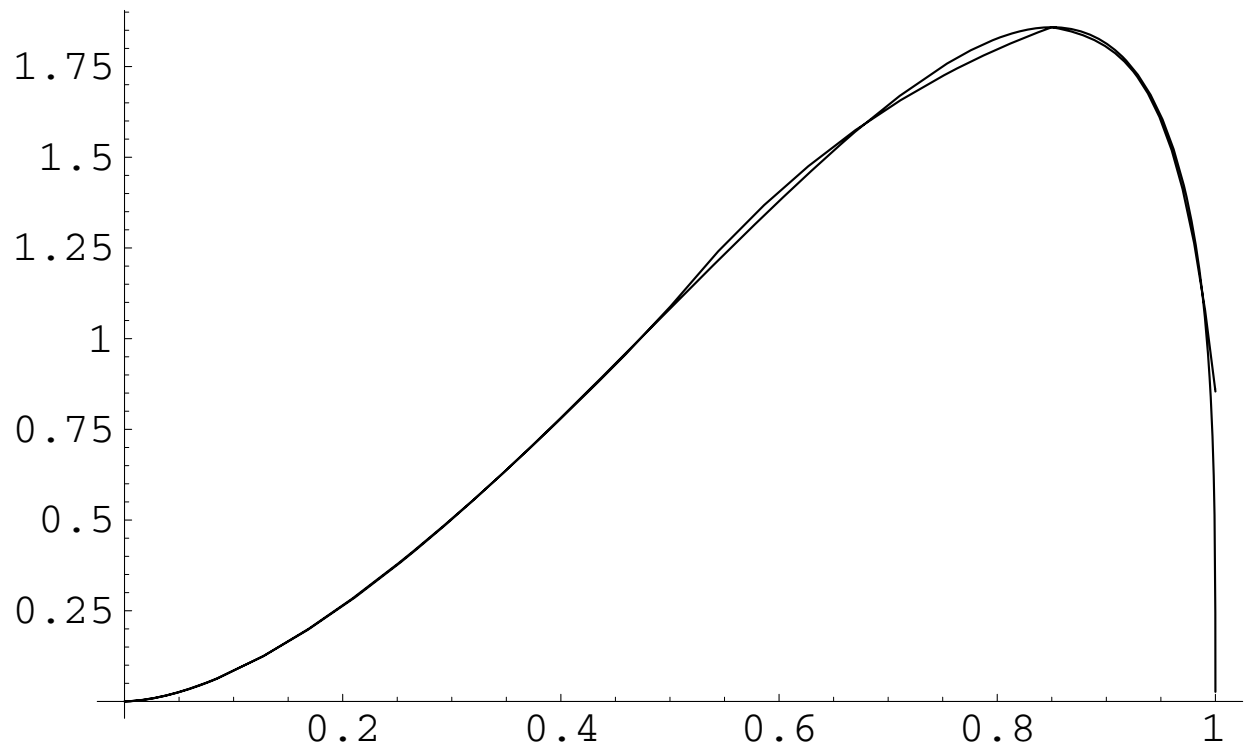


Figure 17: The MTE potential for  $X$  overlaid on the actual  $Beta(2.7, 1.3)$  distribution.

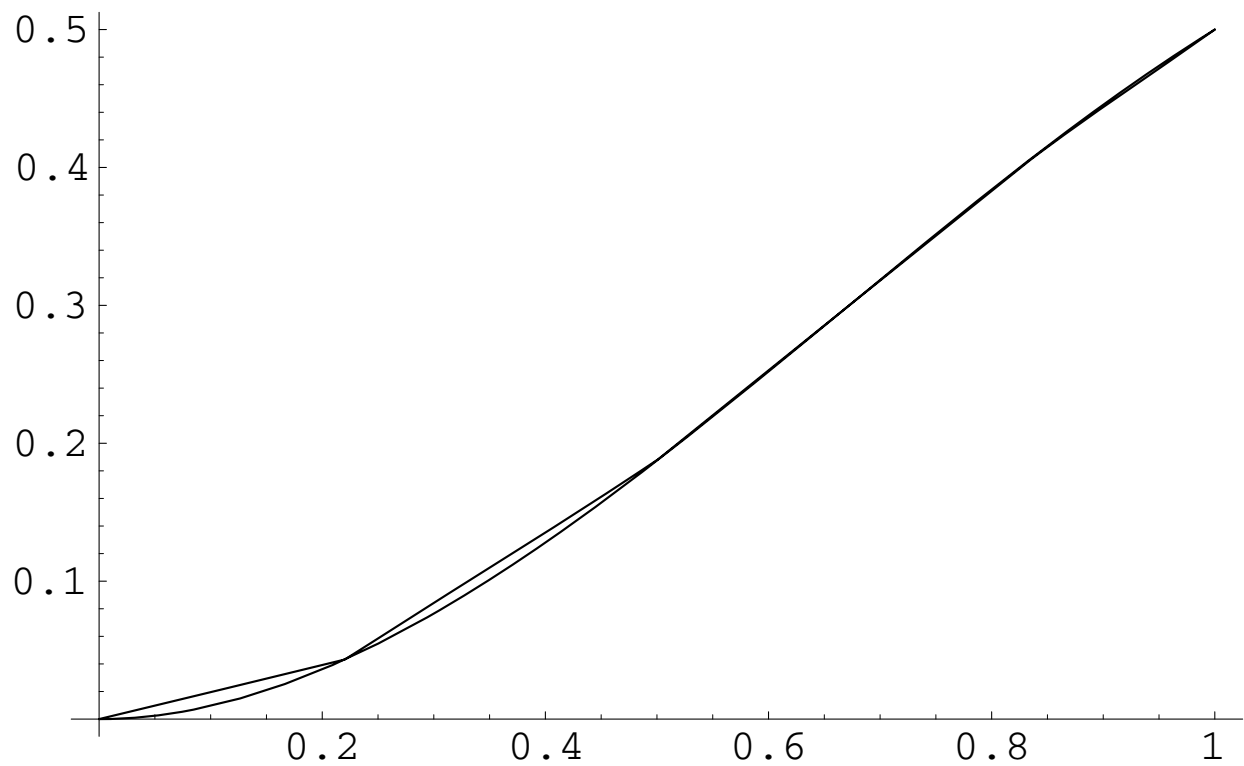


Figure 18: The piecewise linear approximation  $g^{(5)}(x)$  overlaid on the function  $g(x)$  in Example Three.

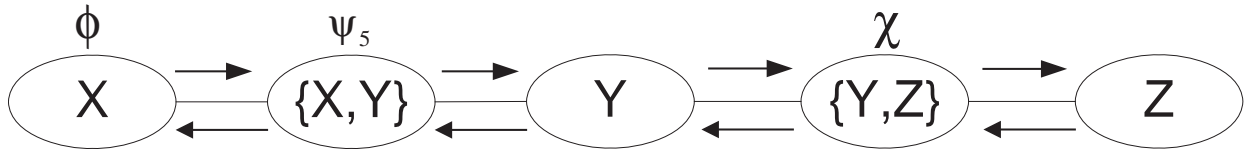


Figure 19: The join tree for the example problem.

The conditional distribution for  $Y$  given  $X$  is represented by a CMF as follows:

$$\psi^{(5)}(x, y) = p_{Y|x}(y) = \mathbf{1}\{y = g^{(5)}(x)\} .$$

The probability distribution for  $Z$  is defined as  $\mathcal{L}(Z | y) \sim N(2y + 1, 1)$  and is approximated by  $\chi$ , which is a two-piece, three-term MTE approximation to the normal distribution [Cobb and Shenoy 2003].

## 5.4 Computing Messages

The join tree for the example problem is shown in Figure 19.

The messages required to calculate posterior marginals for each variable in the network without evidence are as follows:

- 1)  $\phi$  from  $\{X\}$  to  $\{X, Y\}$
- 2)  $(\phi \otimes \psi^{(5)})^{\downarrow Y}$  from  $\{X, Y\}$  to  $\{Y\}$  and  $\{Y\}$  to  $\{Y, Z\}$
- 3)  $((\phi \otimes \psi^{(5)})^{\downarrow Y} \otimes \chi)^{\downarrow Z}$  from  $\{Y, Z\}$  to  $\{Z\}$

## 5.5 Posterior Marginals

The posterior marginal distribution for  $Y$  is the message sent from  $\{X, Y\}$  to  $\{Y\}$  and is calculated using the operation in (9). The expected value and variance of this distribution are calculated as 0.3042 and 0.0159, respectively. The posterior marginal distribution for  $Z$  is the message sent from  $\{Y, Z\}$  to  $\{Z\}$  and is calculated by point-wise multiplication of MTE functions, followed by marginalization (see operations defined in [Cobb and Shenoy 2003]). The expected value and variance of this distribution are calculated as 1.6084 and 1.0455, respectively. The posterior marginal CDF's for  $Y$  and  $Z$  are shown graphically in Figure 20.

## 5.6 Entering Evidence

Suppose we observe evidence that  $Z = 0$  and let  $e_Z$  denote this evidence. Define  $\varphi = (\phi \otimes \psi^{(5)})^{\downarrow Y}$  and  $\psi'^{(5)}(x, y) = \mathbf{1}\{x = g^{(5)-1}(y)\}$  as the potentials resulting from the reversal of the arc between  $X$  and  $Y$  [Cobb and Shenoy 2004]. The evidence  $e_Z$  is passed from  $\{Z\}$  to

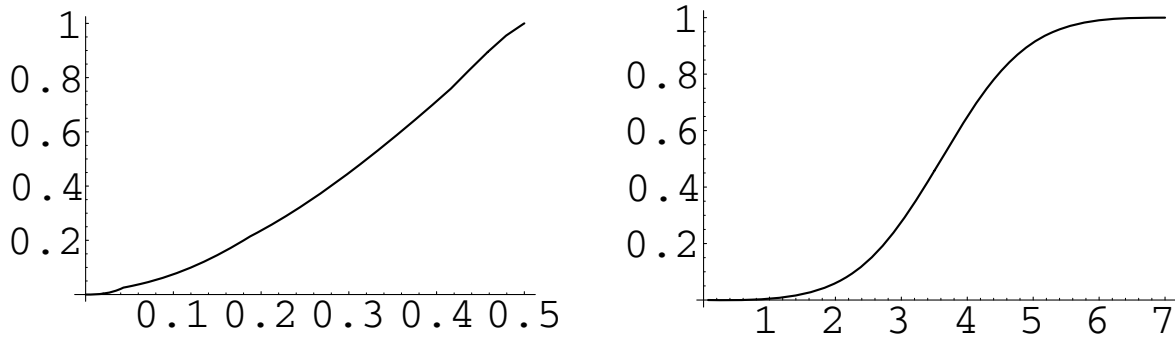


Figure 20: The posterior marginal CDF's for  $Y$  (left) and  $Z$  (right).

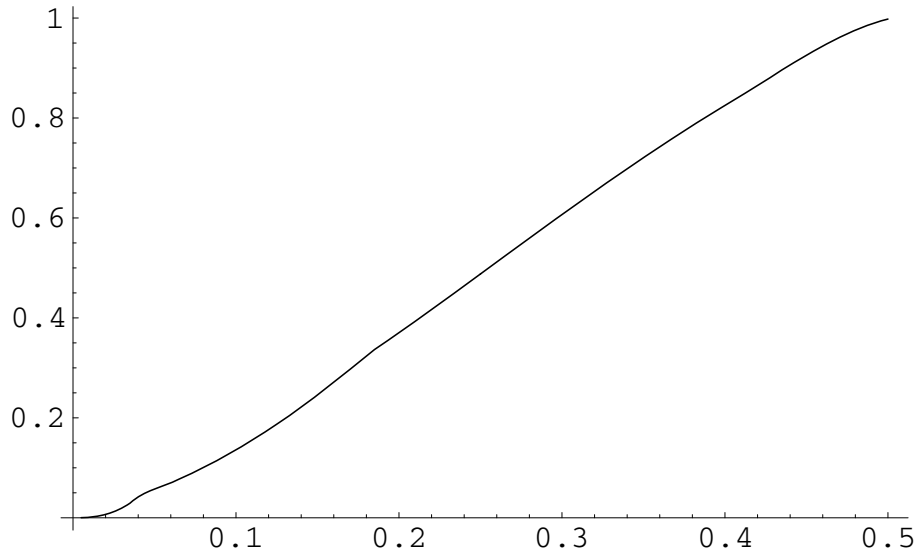


Figure 21: The posterior marginal CDF for  $Y$  considering the evidence  $Z = 0$ .

$\{Y, Z\}$  in the join tree, where the existing potential is restricted to  $\chi(y, 0)$ . This likelihood potential is passed from  $\{Y, Z\}$  to  $\{Y\}$  in the join tree.

Denote the unnormalized posterior marginal distribution for  $B$  as  $\xi'(y) = \varphi(y) \cdot \chi(y, 0)$ . The normalization constant is calculated as  $K = \int_y (\varphi(y) \cdot \chi(y, 0)) dy = 0.0670$ . Thus, the normalized marginal distribution for  $Y$  is found as  $\xi(y) = K^{-1} \cdot \xi'(y)$ . The expected value and variance of this distribution (whose CDF is displayed in Figure 21) are calculated as 0.2560 and 0.0167, respectively.

Using the operation in (9), we determine the posterior marginal distribution for  $X$  as  $\vartheta = (\xi \otimes \psi^{(5)}) \downarrow^X$ . The expected value and variance of this distribution are calculated as 0.5942 and 0.0480, respectively. The posterior marginal distribution for  $X$  considering the evidence is shown graphically in Figure 22. The CDF for the marginal distribution for  $X$  considering the evidence is shown in Figure 23.

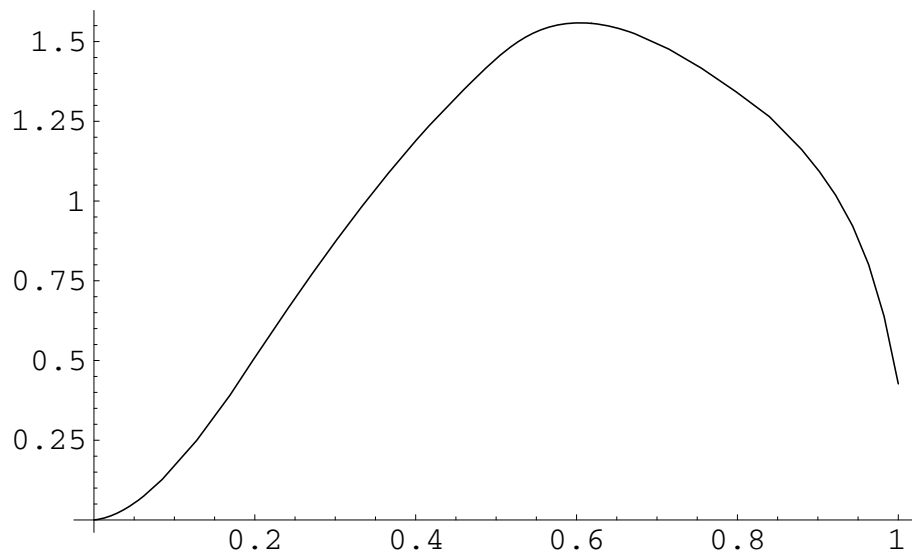


Figure 22: The posterior marginal distribution for  $X$  considering the evidence ( $Z = 0$ ).

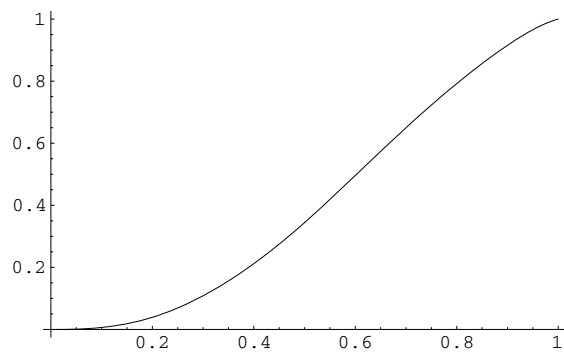


Figure 23: The posterior marginal CDF for  $X$  considering the evidence ( $Z = 0$ ).

## 6 Summary and Conclusions

This paper has described operations required for inference in Bayesian networks containing variables that are nonlinear deterministic functions of their continuous parents. Since the joint PDF for a network with deterministic variables does not exist, the operations required are based on the method of convolutions from probability theory. By approximating nonlinear functions with piecewise linear approximations, we ensure the class of MTE potentials are closed under these operations. Bayesian networks in this paper contain only continuous variables. In future work, we plan to design a general inference algorithm for Bayesian networks that contain a mixture of discrete and continuous variables, with some continuous variables defined as deterministic functions of their continuous parents.

## References

- [1] Cobb, B.R. and P.P. Shenoy (2003), “Inference in hybrid Bayesian networks with mixtures of truncated exponentials,” School of Business Working Paper No. 294, University of Kansas, Lawrence, Kansas. Available for download at: <http://www.people.ku.edu/~brcobb/WP294.pdf>
- [2] Cobb, B.R. and P.P. Shenoy (2004), “Inference in hybrid Bayesian networks with deterministic variables,” in P. Lucas (ed.), *Proceedings of the Second European Workshop on Probabilistic Graphical Models (PGM-04)*, 57–64, Leiden, Netherlands.
- [3] Cobb, B.R., Shenoy, P.P. and R. Rumí (2003), “Approximating probability density functions with mixtures of truncated exponentials,” Working Paper No. 303, School of Business, University of Kansas, Lawrence, Kansas. Available for download at: <http://www.people.ku.edu/~brcobb/WP303.pdf>
- [4] Kullback, S. and R.A. Leibler (1951), “On information and sufficiency,” *Annals of Mathematical Statistics*, 22, 79–86.
- [5] Larsen, R.J. and M.L. Marx (2001), *An Introduction to Mathematical Statistics and its Applications*, Prentice Hall, Upper Saddle River, N.J.
- [6] S.L. Lauritzen and F. Jensen (2001), “Stable local computation with conditional Gaussian distributions,” *Statistics and Computing*, 11, 191–203.
- [7] MacKay, D.J.C. (2003), *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, Cambridge, United Kingdom.
- [8] Moral, S., Rumí, R. and A. Salmerón (2001), “Mixtures of truncated exponentials in hybrid Bayesian networks,” in P. Besnard and S. Benferhart (eds.), *Symbolic and Quantitative Approaches to Reasoning under Uncertainty*, Lecture Notes in Artificial Intelligence, 2143, 156–167, Springer-Verlag, Heidelberg.

- [9] Moral, S., Rumí, R. and A. Salmerón (2002), “Estimating mixtures of truncated exponentials from data,” in J.A. Gamez and A. Salmerón (eds.), *Proceedings of the First European Workshop on Probabilistic Graphical Models (PGM-02)*, 135–143, Cuenca, Spain.
- [10] Moral, S., Rumí, R. and A. Salmerón (2003), “Approximating conditional MTE distributions by means of mixed trees,” in T.D. Nielsen and N.L. Zhang (eds.), *Symbolic and Quantitative Approaches to Reasoning under Uncertainty*, Lecture Notes in Artificial Intelligence, 2711, 173–183, Springer-Verlag, Heidelberg.
- [11] Rumí, R. (2003), *Modelos De Redes Bayesianas Con Variables Discretas Y Continuas*, Doctoral Thesis, Universidad de Almería, Departamento de Estadística Y Matemática Aplicada, Almería, Spain.