Approximate Inference in Credal Networks by Variational Mean Field Methods

Jaime Shinsuke Ide and **Fabio Gagliardi Cozman**
Escola Politécnica, University of São Paulo
Av. Prof. Mello Moraes, 2231 - São Paulo, SP, Brazil
jaime.ide@poli.usp.br, fgcozman@usp.br

Abstract

Graph-theoretical representations for sets of probability measures (credal networks) generally display high complexity, and approximate inference seems to be a natural solution for large networks. This paper introduces a variational approach to approximate inference in credal networks: we show how to formulate mean field approximations using naive (fully factorized) and structured (tree-like) schemes. We discuss the computational advantages of the variational approach, and present examples that illustrate the mechanics of the proposal.

Keywords. Credal networks, variational methods, inferences.

1 Introduction

Graphical models that represent uncertainty through sets of probability measures are often referred to as *credal networks* [2, 8]. In a credal network, a collection of sets of probability measures is associated with a directed acyclic graph. *Inference* in a credal network usually means the computation of lower and upper bounds for the conditional probability of an event. The complexity of inference in credal networks is generally high (even for tree-like networks [5]), and approximate inference seems to be a natural solution for large networks [1].

In this work we present a new approach for approximate inference in credal networks — we propose a variational approach. Modern variational methods are popular in various fields such as control theory, optimization, statistics, economics and machine learning; recently several variational approaches have been successfully used for inference and estimation of densely connected graphical probability models [15, 16]. To the best of our knowledge, we are the first to explore an explicit variational formulation for inference in credal networks. There has been previous work on approximations that minimize Kullback-Leibler distance, such as the work by Cano and Moral [1]; these efforts can be viewed as special cases (where a particular structure is used in the approximation) of a broader variational formulation.

Variational methods are rather general, as they can be used to handle categorical and continuous distributions. In this paper we pursue a somewhat restricted formulation, as a first step in understanding the flexibility of the approach.

In Sections 2 and 3 we give a brief review of credal networks and variational methods. Our main contribution, a variational approach for inference in credal networks, is presented in Section 4. Implementation details and some empirical results are described in Sections 5 and 6.

2 Credal Networks

In this section we present a few facts on credal networks (and their basic elements, *credal sets* [11]); a more detailed discussion can be found elsewhere [2, 4, 8].

A credal set for variable X is denoted by $K(X)$; we assume that every variable is categorical and that every credal set is convex, closed, and has a finite number of vertices. Given a credal set $K(X)$ and a function $f(X)$, the *lower* and *upper* expectations of $f(X)$ are defined respectively as $\underline{E}[f(X)] = \min_{p(X)\in K(X)} E_p[f(X)]$ and $\overline{E}[f(X)] = \max_{p(X)\in K(X)} E_p[f(X)]$ (here $E_p[f(X)]$ indicates standard expectation). The *lower probability* and the *upper probability* of event E are defined respectively as $\underline{P}(E) = \min_{p(X)\in K(X)} P(E)$ and $\overline{P}(E) = \max_{p(X)\in K(X)} P(E)$. A *conditional credal set* is a set of conditional distributions, obtained by applying Bayes rule to each distribution in a credal set of joint distributions [18]. Lower and upper conditional probabilities for a variable X given an event E are defined accordingly:

$\underline{P}(X=x|E) \;=\; \min_{p(X)\in K(X)}(P(X=x,E)\,/P(E))$, $\overline{P}(X=x|E) = \max_{p(X)\in K(X)}(P(X=x,E)\,/P(E))$.

Credal networks are directed acyclic graphs associated with credal sets. An *inference* is the computation of lower and upper probabilities for an event $\{X_q = x_q\}$ given *evidence* D — here D indicates a set of observed variables. Those variables that do not belong to D are called *hidden* variables and are denoted by H. A credal network is defined by local *separately specified* credal sets $\{K(X|Y=y)\}$ when these credal sets are not related for different values of the conditioning variables Y. The *strong extension* of the network is the convex hull of the set containing all joint distributions that factorize as $\prod_i p(X_i|pa(X_i))$, where each conditional distribution $p(X_i|pa(X_i)=\pi_k)$ is selected from the local credal set $K(X_i|pa(X_i)=\pi_k)$ [3]. A strong extension defines a joint credal set where each vertex is a possible combination of conditional distributions (each vertex of this set can be represented by a Bayesian network). There is a vertex that minimizes and a vertex that maximizes $P(X=x|E)$ [8]; thus an inference in a strong extension can be viewed as an optimization problem over the set of potential vertices.

Exact inference in general credal networks displays high complexity. Apparently the only tractable situation is represented by polytree-shaped networks with binary variables [8]. Other than this, even inference in polytree-shaped credal networks is a NP-complete problem [5]. The most promising methods for exact inference are based on multilinear programming [7], but even these methods face difficulties for networks of medium size (say twenty to twenty five reasonably connected nodes). Due to the complexity of exact inference, several algorithms for approximate inference have been developed [1, 6, 9, 17].

3 Variational Methods and Mean Field Approximations

In this section we give a brief review of variational methods, focusing on those methods that are applied to Bayesian networks. We follow the terminology used in previous introductory material [10, 19].

3.1 Basic notions

Suppose that we have a directed graph associated with a joint distribution $P(X)$, where X represents the set of variables. We want to approximate $P(H|D)$ by a distribution $Q(H)$, where H is the set of hidden variables and D is the evidence — thus $X = \{H, D\}$. We choose the Kullback-Leibler (KL) divergence as a dissimilarity measure between $P(H|D)$ and $Q(H)$:

$$\begin{aligned} KL(Q||P) &= \sum_H Q(H)\ln\frac{Q(H)}{P(H|D)} \\ &= \sum_H Q(H)\ln\frac{Q(H)}{P(H,D)} + \ln P(D) \\ &= \sum_H Q(H)\ln Q(H) + \ln P(D) \\ &\quad -\sum_H Q(H)\ln P(H,D) \qquad (1) \end{aligned}$$

The first term of the last expression is the negative entropy [10]. The second term $\ln P(D)$ is constant with respect to $Q(H)$. The last term is the expectation of $P(H,D)$ with respect to $Q(H)$. From Equation (1), the divergence is null (KL = 0) when $Q(H) = P(H|D)$. The goal here is to find a good approximation $Q(H)$ to $P(H|D)$ by minimizing $KL(Q||P)$.

3.2 The naive mean field approximation

An approximate model that has been successfully used in variational methods in many areas is the fully factorized distribution; this is often called the *mean field* approximation [13]. The idea is that the global behavior of distributions should be approximated by a set of independent variables [15]. Using the fully factorized distribution, we can minimize the divergence KL in an iterative and computationally efficient manner.

Consider then a fully factorized distribution $Q(H) = \prod_i Q_i(H_i)$. Substituting in Equation (1) we obtain:

$$\begin{aligned} KL(Q||P) &= \sum_H \prod_i Q_i(H_i)\ln(\prod_i Q_i(H_i)) \\ &\quad -\sum_H \prod_i Q_i(H_i)\ln P(H,D) + \ln P(D) \\ &= \sum_i \sum_{H_i} Q_i(H_i)\ln Q(H_i) \\ &\quad -\sum_H \prod_i Q_i(H_i)\ln P(H,D) + \ln P(D) \\ &= -\sum_i \mathbf{H}(Q_i) + \ln P(D) \\ &\quad -\sum_H \prod_i Q_i(H_i)\ln P(H,D), \qquad (2) \end{aligned}$$

where $\mathbf{H}$ represents the entropy. The idea is to minimize KL with respect to $Q_j(H_j)$ by assuming fixed terms $Q_i(H_i)$, $i \neq j$. This is done by first separating out the Equation (2) in terms of $Q_j(H_j)$:

$$KL(Q||P) = -\mathbf{H}(Q_j) - \sum_{i\neq j}\mathbf{H}(Q_i)$$

$$-\sum_{H_j} Q_j(H_j) \sum_{H_{i\neq j}} \prod_i Q_i(H_i) \ln P(H,D) + \ln P(D),$$

and differentiating this expression with respect to $Q_j(H_j)$. Note that we must take into account the normalization constraint $\sum_{h_j \in H_j} Q(h_j) = 1$, where h_j is a value of H_j. This is done by introducing the Lagrange parameter λ and then solving the equation:

$$\frac{\partial}{\partial Q_j(h_j)}[KL(Q||P) - \lambda(\sum_{h_j \in H_j} Q(h_j) - 1)] = 0. \quad (3)$$

Then we obtain the expression of $Q_j^*(h_j)$ that partially minimizes $KL(Q||P)$:

$$\ln Q_j^*(h_j) = \sum_{H_{i\neq j}} \prod_i Q_i(H_i) \ln P(H,D) + \lambda - 1. \quad (4)$$

The normalization constant $k_\lambda = \exp(1-\lambda)^{-1}$ is calculated computing Equation (4) for all values of the variable H_j:

$$k_\lambda = \sum_{h_j \in H_j} Q_j^*(h_j).$$

Equation (4) is the unique solution of Expression (3). Hence this is the global minimum of $KL(Q||P)$ with respect to $Q_j(H_j)$. Once we update variable H_j, we must choose another variable H_i to be updated, and so on for all hidden variables H. This process minimizes the KL divergence iteratively.

3.3 The importance of local computation

Consider the computational cost of the updating Equation (4). The key point here is that this updating expression can be simplified so that only "local" computations are needed for each hidden variable H_j ("local" in the sense that they only refer to terms that are "close" in the graph underlying the network).

Consider first that the joint distribution $P(H,D)$ can be written in terms of conditional distributions. Take Equation (4) and substitute $P(H,D)$ by local conditional distributions $P(X_k|pa_k)$, where pa_k is the set of parents of X_k:

$$\begin{aligned} Q_j^*(H_j) &= k_\lambda \exp\left(\sum_{H_{i\neq j}} \prod_i Q_i(H_i) \ln P(H,D)\right) \\ &= k_\lambda \exp\left(\sum_{H_{i\neq j}} \prod_i Q_i(H_i) \sum_k \ln P(X_k|pa_k)\right). \end{aligned}$$

Now, every conditional distribution $P(X_k|pa_k)$ that does not depend on variable H_j will result in a term that is constant over H_j (and is summed out). We obtain:

$$\begin{aligned} Q_j^*(H_j) = k_\lambda \exp(&\sum_{H_{i\neq j}} \prod_i Q_i(H_i)[\ln P(H_j|pa_j) \\ &+ \sum_{k \in ch_j} \ln P(X_k|pa_k)]). \end{aligned}$$

As every term $Q_i(H_i)$ that does not belong to $\{pa_j\}$ and $\{X_k, pa_k\}$ will be summed out, we obtain:

$$\begin{aligned} Q_j^*(H_j) \quad = \quad & k_\lambda \exp(\sum_{H_i \in \{pa_j\}} \prod_i Q_i(H_i) \ln P(H_j|pa_j) \\ &+ \sum_{k \in ch_j} \sum_{H_{i\neq j} \in \{X_k, pa_k\}} \prod_i Q_i(H_i) \ln P(X_k|pa_k)), \quad (5) \end{aligned}$$

where ch_j is the set of children of H_j. This means that we only have to make "local" computations involving conditional distributions for variables in the Markov blanket of H_j (the *Markov* blanket of variable H_j is the set of nodes containing the parents of H_j, the children of H_j, and the parents of children of H_j).

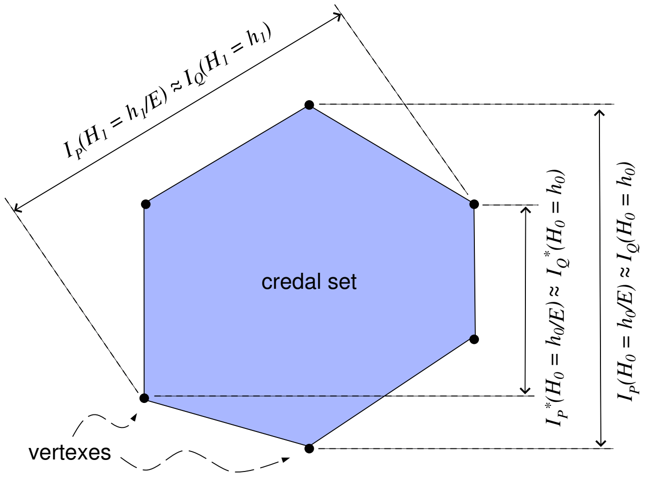
4 A Variational Method for Inference in Credal Networks

We want to derive a variational method for inference in credal networks. A direct approach is a vertex-based approximation: Apply a mean field method in each vertex of the joint distribution; that is, approximate individually all vertices of the strong extension of local credal sets $K_P(H_j|pa_j)$. Such a vertex-based approach does not really solve the main problem with strong extensions — the enormous growth in the number of potential vertices.

A second approach is to focus on a set-based approximation. The goal here is to approximate the original joint credal set by a set of probability intervals $I_Q(H_j = h_j) = [\underline{Q}(H_j = h_j), \overline{Q}(H_j = h_j)]$ for each value of each variable H_j. These intervals are directly related to the lower and upper probabilities $\underline{P}(H_j = h_j|E)$ and $\overline{P}(H_j = h_j|E)$. The set-based approach mimics the local computations in the "standard" mean field methods, but the local computations are replaced by interval computations — the result is a process that iteratively computes probability intervals for all variables H_j.

Consider first the updating Equation (5) and upper/lower bounds $\underline{Q}(H_j = h_j)$ and $\overline{Q}(H_j = h_j)$:

$$\underline{Q}_j^*(H_j = h_j) \quad \propto$$



$I_p(H_1 = h_1/E) \approx I_Q(H_1 = h_1)$
credal set
vertexes
$I_P{}^*(H_0 = h_0/E) \approx I_Q{}^*(H_0 = h_0)$
$I_p(H_0 = h_0/E) \approx I_Q(H_0 = h_0)$

Figure 1: The outer bound step. Interval $I^*_Q(H_0 = h_0)$ is approximated by an outer bound interval $I_Q(H_0 = h_0)$.

$$\min \sum_{H_i \in \{pa_j\}} \prod_i Q_i(H_i) \ln P(H_j|pa_j) + \sum_{k \in ch_j} \sum_{H_{i \neq j} \in \{X_k, pa_k\}} \prod_i Q_i(H_i) \ln P(X_k|pa_k), (6)$$

$$\overline{Q}^*_j(H_j = h_j) \quad \propto \quad \max \sum_{H_i \in \{pa_j\}} \prod_i Q_i(H_i) \ln P(H_j|pa_j) + \sum_{k \in ch_j} \sum_{H_{i \neq j} \in \{X_k, pa_k\}} \prod_i Q_i(H_i) \ln P(X_k|pa_k), (7)$$

where:

$$\begin{aligned} Q_i(H_i) &\in I_Q(H_i = h_i), \\ P(H_j|pa_j = \pi_j) &\in K_P(H_j|pa_j = \pi_j), \\ P(X_k|pa_k = \pi_k) &\in K_P(X_k|pa_k = \pi_k). \end{aligned}$$

Unfortunately, the exact computation of the intervals $I_Q(H_j = h_j)$ leads us to a global combinatorial search, because these intervals are all interrelated. A vertex that attains the minimum or maximum of $Q(H_j = h_j)$ is not necessarily the same vertex that attains the minimum or maximum of $Q(H_i = h_i)$ for $i \neq j$; enforcing such relationships leads us to combinatorial explosion.

We now consider an approximate solution that circumvents the complexity problem just discussed. In fact, we introduce a second approximation on top of the variational one. The idea is to make intervals $I_Q(H_j = h_j)$ unrelated to each other, thus reducing the computation of an approximate interval to a truly "local" combinatorial problem. To do this, we update $I_Q(H_j = h_j)$ by Equations (6) and (7), using the outer bounds of $I_Q(H_i = h_i)$ for $i \neq j$. That is, we use approximate values $I_Q(H_i = h_i)$, computed for those vertices that $Q(H_j = h_j)$ is maximum and minimum, instead of vertices where $Q(H_i = h_i)$ is maximum or minimum. We call this approximation an *Outer Bound Step.*

The outer bound step is depicted in Figure 1. Suppose that we are approximating $I_P(H_0, H_1)$ by $I_Q(H_0, H_1)$, where H_0 and H_1 are hidden variables, using a variational method. To update the interval $I_Q(H_1 = h_1)$ [1], we need the values of $I^*_Q(H_0 = h_0)$ for those vertices where $I_Q(H_1 = h_1)$ are attained. For variable H_0 we just consider those vertices that lead to the extreme values of $I_Q(H_0 = h_0)$ — we take these values as an outer bound approximation for interval $I^*_Q(H_0 = h_0)$. Note that such a round of approximations happens in each iteration of the mean field scheme. The method can be summarized as follows:

- **Mean field approximation.** We approximate the original strong extension $K_P(X)$ by an approximate credal set $K_Q(\mathbf{X})$ that is the strong extension of local probability intervals $I_Q(H_j = h_j) = [\underline{Q}(H_j = h_j), \overline{Q}(H_j = h_j)]$ for each value of $Q(H_j)$. Here $\underline{Q}(H_j = h_j)$ approximates the minimum value of $P(H_j = h_j|E)$ and $\overline{Q}(H_j = h_j)$ approximates the maximum value. This approximation moves us to a local computation.

- **Outer bound step.** We update the interval $I_Q(H_j = h_j)$ by Equations (6) and (7), using the outer bounds $I_Q(H_i = h_i)$, for $i \neq j$. This approximation step is crucial to keep the local computation property sought by the variational approximation.

We transform the global combinatorial problem into a local one, because: (1) The mean field approximation is based on a "local" updating mechanism (Equation (5)), restricted to the Markov Blanket; (2) The outer bound step makes each updating step combinatorially unrelated to all others, thus guaranteeing locality of computation.

Figure 2 presents an algorithm that implements a variational mean field approach for inference in a credal network.

[1] Notice that $I_Q(H_1 = h_1)$ is a approximation of $I_P(H_1 = h_1|E)$, and so on.

```
Set-based Mean Field
            Input: Credal network $N$.
            Output: Approximations for lower/upper probabilities of some variable $H_j$.
01. Initialize the intervals $I_Q(H = h) = [\underline{Q}^*(H), \overline{Q}^*(H)]$, for all values of hidden variable $H$.
02. For each hidden variable $H_i$, repeat until convergence:
03.     For each possible vertex in the Markov Blanket of $H_i$ (Equations 6 and 7):
04.         Compute $Q^*(H_i)$ for each value of $H_i$ and normalize $Q^*(H_i = h_i)$.
05.         Keep the minimum and maximum value of $Q^*(H_i = h_i)$ and update $I_Q(H_i = h_i)$ .
06. Return $\underline{Q}(H_i = h_i)$ and $\overline{Q}(H_i = h_i)$ as approximations for $\underline{P}(H_i = h_i|E)$ and $\overline{P}(H_i = h_i|E)$.
```

Figure 2: Mean field in credal networks.

4.1 Structured variational approximation: the binary case

The naive variational mean field approach described previously is computationally attractive, but it is often unable to yield sufficiently accurate results [10]. A natural idea to improve over the naive mean field method is to combine it with exact calculations — for example, to approximate the original intractable network with tractable substructures such as trees and chains [16]. We introduce a *Structured Variational-2U Algorithm* (SV2U) that uses the set-based variational approach to approximate a multi-connected binary credal network by a polytree-structure network and run the exact interval propagation *2U algorithm* [8]. Note that for binary variables the intervals $I_Q = [p_{low}, p_{high}]$ (of Equation (6)) corresponds to the local credal sets K_Q.

Consider the idea of *structured* variational method [10, page 18]. We take a factor Q_i over a cluster of variables $c_i = \{X_i, pa_i\}$, so as to build the approximate joint distribution: $Q(X) = \prod_i Q_i(c_i)$. We have a more complex updating equation than in the naive mean field method, because during the minimization of $KL(Q||P)$, some entropy terms are no longer constants. We obtain the updating equation for the cluster $c_j = \{X_j, pa_j\}$ [19, page 104]:

$$\begin{aligned} Q_j^*(X_j|pa_j) \quad &\propto \\ \sum_{k \in G_{X_j}} \sum_{l \in \{\mathbf{X}/c_j\}} &\prod_l Q_l(X_l) \ln P(X_k|pa_k) \\ - \sum_{i \in C_{X_j}} \sum_{l \in \{\mathbf{X} \backslash c_j\}} &\prod_l Q_l(X_l) \ln Q_i(X_i|pa_i), \quad (8) \end{aligned}$$

where G_{X_j} represents the set of clusters $g_k = \{X_k, pa_k\}$ that depend on X_j in the original network and C_{X_j} is the set of clusters c_i that depend on X_j in the approximate network, excluding c_j itself (in Bayesian networks, these sets of clusters are Markov blankets). Note that the expectation in Equation (8), $E\Big[\sum_{k \in G_{X_j}} \ln P(X_k|pa_k) - \sum_{i \in C_{X_j}} \ln Q_i(X_i|pa_i)\Big]$, is computed with respect to the approximated structure $Q(X) = \prod_i Q_i(c_i)$, $i \neq j$. Also note that we need to update just those clusters c_j that do not belong to the original network cluster set G; that is, we have to update just those conditional distributions that are modified.

We use this structured variational approach described before and develop it for credal networks. Apply the set-based approach to the Equation (8), and obtain the updating expressions analogous to Equation (6):

$$\begin{aligned} \underline{Q}_j^*(X_j|pa_j) \quad &\propto \\ \min \sum_{k \in G_{X_j}} \sum_{l \in \{\mathbf{X} \backslash c_j\}} &\prod_l Q_l(X_l|pa_l) \ln P(X_k|pa_k) \\ - \sum_{i \in C_{X_j}} \sum_{l \in \{\mathbf{X} \backslash c_j\}} &\prod_l Q_l(X_l|pa_l) \ln Q_i(X_i|pa_i), (9) \end{aligned}$$

$$\begin{aligned} \overline{Q}_j^*(X_j|pa_j) \quad &\propto \\ \max \sum_{k \in G_{X_j}} \sum_{l \in \{\mathbf{X} \backslash c_j\}} &\prod_l Q_l(X_l|pa_l) \ln P(X_k|pa_k) \\ - \sum_{i \in C_{X_j}} \sum_{l \in \{\mathbf{X} \backslash c_j\}} &\prod_l Q_l(X_l|pa_l) \ln Q_i(X_i|pa_i), (10) \end{aligned}$$

where:

$$\begin{aligned} Q_l(X_l|pa_l = \pi_l) &\in K_Q(X_l|pa_l = \pi_l), \\ Q_i(X_i|pa_i = \pi_i) &\in K_Q(X_i|pa_i = \pi_i), \\ P(X_k|pa_k = \pi_k) &\in K_P(X_k|pa_k = \pi_k). \end{aligned}$$

The proposed set-based structured variational algorithm (SV2U) is described in Figure 3 for credal networks containing binary variables. Suppose that we have a multi-connected binary network N with joint distribution $P(X) = \prod_k P_k(g_k)$. The idea is to approximate this with a polytree-structure network N_p, finding a *loopy cutset* [14], with distribution $Q(X) = \prod_i Q_i(c_i)$ (Line 01). Define the set of clusters $G = \{g_k\}$ of network N and $C = \{c_i\}$

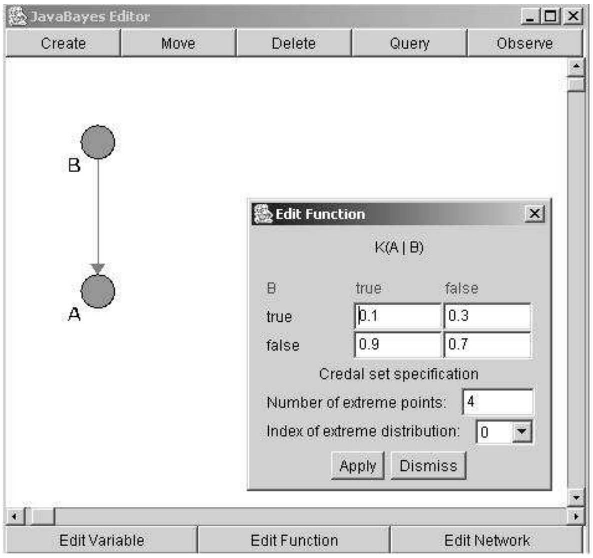
```
Structured Variational 2U: SV2U
        Input: Multi-connected binary credal network $N$.
        Output: Approximations for lower/upper probabilities of some variable $X_j$.
01. Find a cutset of the original network $N$,with joint distribution $P(X) = \prod_k P_k(g_k)$, to obtain
    an approximate polytree structure $N_p$.
02. Find the set of clusters $G = \{g_k\}$ of network $N$ and $C = \{c_i\}$ of the network $N_p$.
03. Initialize for those clusters $c_i$ that don't belong to $G$, $c_j$, $K_Q(X_j = x_j|pa_j = \pi_j) = [0.5, 0.5]$.
03. For those clusters $c_j$ repeat until convergence:
04.     For all possible parent configurations of cluster $c_j$, $\{pa_j = \pi_j\}$:
05.         Compute $\underline{Q}^*(X_j = x_j|pa_j = \pi_j)$ and $\overline{Q}^*(X_j = x_j|pa_j = \pi_j)$ for each
            value of $X_j$ and normalize these values, from Equations (9) and (10).
06.         Keep the minimum and maximum values of $Q^*(X_j = x_j|pa_j = \pi_j)$ and
            update $K_Q(X_j = x_j|pa_j = \pi_j)$ .
07. Run the 2U algorithm in the polytree-structure associated to the joint distribution $Q(X) =
    \prod_k P_k(g_k) \times \prod_j Q_j(c_j)$, where $k$ is the index of those clusters that belong to both cluster sets $G$
    and $Q_j(c_j) \in K_Q(X_j = x_j|pa_j = \pi_j)$. Keep the minimum and maximum values of probabilities of
    some variable $X_j$.
```

Figure 3: The SV2U algorithm.

of the network N_p. Find the approximate conditional distributions of those clusters c_i that does not belong to the set G. Define this set as c_j. This means that we have to compute the distributions $Q(c_j)$. This is done iteratively, as described in Section 3.2, until we obtain convergence of these distributions (Lines 03-06). During this iteration, we compute the $K_Q(X_j = x_j|pa_j = \pi_j) \in [\underline{Q}^*(X_j = x_j|pa_j = \pi_j), \overline{Q}^*(X_j = x_j|pa_j = \pi_j)]$. This means that at the end of the process we have a polytree-structure distribution $Q(X) = \prod_k P_k(g_k) \times \prod_j Q_j(c_j)$, where k is the index of those clusters that belong to both cluster sets G and C; and $Q_j(c_j) \in K_Q(X_j = x_j|pa_j = \pi_j)$. With this polytree-structure distribution we compute the desired approximations for lower/upper probabilities, using the exact algorithm 2U (Line 07) [8].

5 Examples

In this section we present examples that illustrate the mechanics of our mean field approach to approximate inference. The two examples are arguably the simplest possible examples of naive and structure mean field methods respectively; the purpose of these examples is only to show the kinds of expressions that appear in the method. We later discuss larger examples; certainly we cannot claim that these experiments show all empirical properties of set-based mean field methods, but they illustrate several interesting characteristics of the methods.



K(A | B)

B	true	false
true	0.1	0.3
false	0.9	0.7

Credal set specification

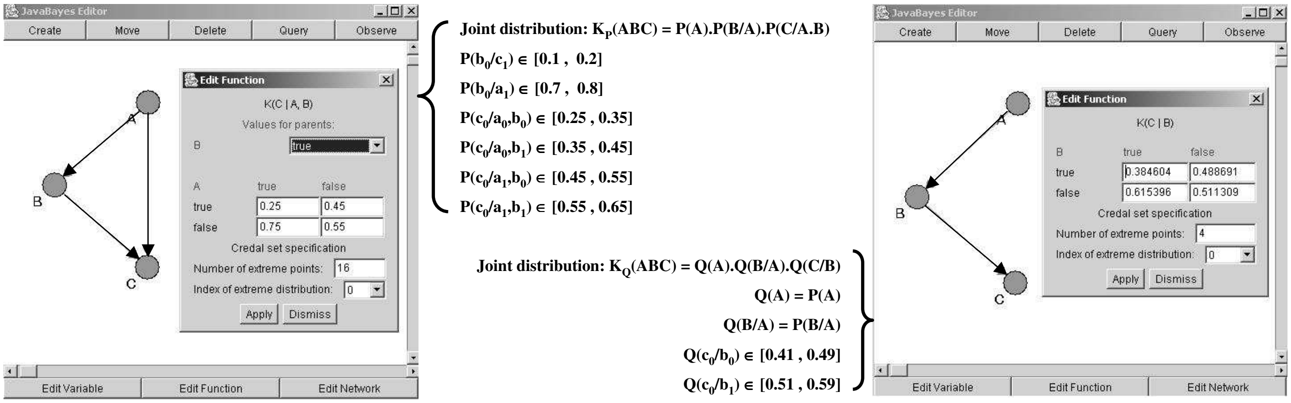
Number of extreme points: 4

Index of extreme distribution: 0

Figure 4: Credal network with 2 binary variables (network in the JavaBayes system, a package that can represent local credal sets and strong extensions).

5.1 Naive mean field

Suppose we have a credal network with 2 binary hidden variables A and B (Figure 4). Consider that it is associated with a credal set defined by intervals (for binary variables, the interval representation is equivalent to the vertex representation): $P(b_0) \in [0.4, 0.75]$, $P(a_0|b_0) \in [0.1, 0.2]$ and $P(a_0|b_1) \in [0.3, 0.4]$. We want to approximate the interval of values for $P(a_0)$.



K(C | A, B)
Values for parents:
B true

A	true	false
true	0.25	0.45
false	0.75	0.55

Credal set specification
Number of extreme points: 16
Index of extreme distribution: 0

Joint distribution: $K_P(ABC) = P(A).P(B/A).P(C/A.B)$
$P(b_0/c_1) \in [0.1\ ,\ 0.2]$
$P(b_0/a_1) \in [0.7\ ,\ 0.8]$
$P(c_0/a_0,b_0) \in [0.25\ ,0.35]$
$P(c_0/a_0,b_1) \in [0.35\ ,0.45]$
$P(c_0/a_1,b_0) \in [0.45\ ,0.55]$
$P(c_0/a_1,b_1) \in [0.55\ ,0.65]$

Joint distribution: $K_Q(ABC) = Q(A).Q(B/A).Q(C/B)$
$Q(A) = P(A)$
$Q(B/A) = P(B/A)$
$Q(c_0/b_0) \in [0.41\ ,0.49]$
$Q(c_0/b_1) \in [0.51\ ,0.59]$

K(C | B)

B	true	false
true	0.384604	0.488691
false	0.615396	0.511309

Credal set specification
Number of extreme points: 4
Index of extreme distribution: 0

Figure 5: Left: multi-connected credal network with 3 binary variables. Right: a structured (polytree-shaped) approximation to the original network, differing in the credal set associated with variable C.

From Equations (6) and (7), we have the updating expressions for variable A:

$$\begin{aligned}\underline{Q}^*(a_0) &= \min k_\lambda.\exp(Q^*(b_0).\ln[P(a_0|b_0)] + \\ &\qquad Q^*(b_1).\ln[P(a_0|b_1)]),\\ \overline{Q}^*(a_0) &= \max k_\lambda.\exp(Q^*(b_0).\ln[P(a_0|b_0)] + \\ &\qquad Q^*(b_1).\ln[P(a_0|b_1)]),\end{aligned}$$

where $Q^*(b_0) \in I(b_0) = [\underline{Q}^*(b_0), \overline{Q}^*(b_0)]$. The normalization constant k_λ is obtained by computing $Q^*(a_1)$ for the same vertex. For variable B:

$$\begin{aligned}\underline{Q}^*(b_0) &= \min k_\lambda.\exp(Q^*(a_0).\ln[P(a_0|b_0).P(b_0)] + \\ &\qquad Q^*(a_1).\ln[P(a_1|b_0).P(b_0)]),\\ \overline{Q}^*(b_0) &= \max k_\lambda.\exp(Q^*(a_0).\ln[P(a_0|b_0).P(b_0)] + \\ &\qquad Q^*(a_1).\ln[P(a_1|b_0).P(b_0)]),\end{aligned}$$

where: $Q^*(a_0) \in I(a_0) = [\underline{Q}^*(a_0), \overline{Q}^*(a_0)]$. The normalization constant k_λ is obtained by computing $Q^*(b_1)$ for the same vertex. These expressions are iteratively updated until convergence. Note that at the beginning the interval $I(b_0)$ is initialized with arbitrary values between 0 and 1. At convergence, we obtain the interval $Q(a_0) \in [0.13, 0.31]$. For comparison, the exact interval is $P(a_0) \in [0.15, 0.32]$.

5.2 Structured mean field

Consider now the structured mean field variational approximation (Section 4.1) to a credal network with 3 binary variables A, B and C. The network is depicted in Figure 5. For each vertex of the credal network, we have a distribution $P(A, B, C) = P(A).P(B|C).P(C|A, B)$ that is approximated by $Q(A, B, C) = P(A).P(B|C).Q(C|B)$ — the original network is approximated by a tree, for which exact inference is available through the 2U algorithm [8], and also approximate inference is possible [5, 9].

In this example, only variable C must be updated. Developing Equation (8), where $G_{X_j} = g_C = \{C, A, B\}$ and $C_{X_j} = c_B = \{B, A\}$, we obtain:

$$\begin{aligned}\underline{Q}^*(C|B) &\propto \min \sum_A P(A)[\sum_{k\in d_C} \ln P(X_k|pa_k) \\ &\quad - \sum_{i\in c_B} \ln Q_i(X_i|pa_i)] \\ &= \min \sum_A P(A)[\ln P(A) + \ln P(B|A) \\ &\quad + \ln P(C|A,B) - \ln P(B|A)] \\ &= \min \sum_A P(A) \ln P(A) \\ &\quad + \sum_A P(A) \ln P(C|A,B) \\ &= \min \sum_A P(A) \ln P(C|A,B),\end{aligned}$$

and analogously:

$$\begin{aligned}\overline{Q}^*(C|B) &\propto \max \sum_A P(A)[\sum_{k\in d_C} \ln P(X_k|pa_k) \\ &\quad - \sum_{i\in c_B} \ln Q_i(X_i/pa_i)] \\ &= \max \sum_A P(A) \ln P(C|A,B).\end{aligned}$$

From these expressions, we get a set of approximate conditional distributions $Q(c_0|b_0) \in [0.41, 0.49]$ and $Q(c_0|b_1) \in [0.51, 0.59]$. Using this approximate credal set, we can infer $Q(c_0) \in [0.4262, 0.5513]$, while the exact result of the inference is $P(c_0) \in [0.428, 0.552]$.

6 Tests and Results

We have implemented the set-based mean field algorithm (Figure 2) and conducted several tests. Accu-

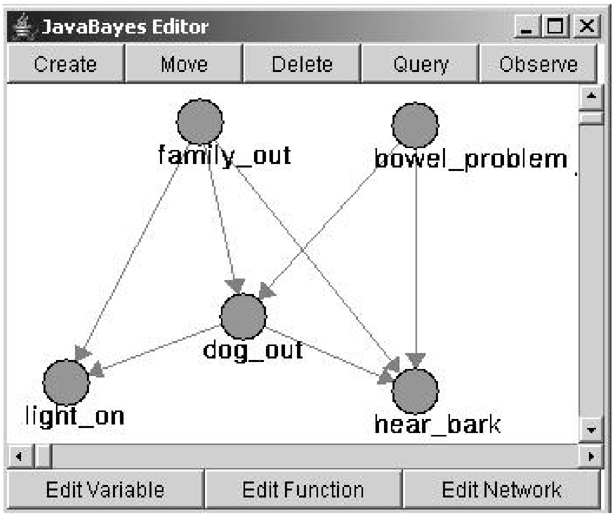


Figure 6: Multidog network. A credal network with 5 binary variables.

racy of the naive method is reasonable but far from spectacular. For example, in the Multidog network (Figure 6), comparing the approximated results obtained by set-based mean field and exact results, we get a mean square error of 7%. Tests with the SV2U indicate a much better performance of the structured mean field approach.

To illustrate the performance of SV2U algorithm, consider a medium sized example. The *Pyramid* network (Figure 7) is a multilayered graph associated with 28 binary variables (assume values ”0” or ”1”) and local connections among layers. [12]. We associate each variable with a binary credal set, a convex hull of the set containing all joint distributions that factorize as $\prod_i P(X_i|pa(X_i))$, $i = \{1, ..., 28\}$, where each conditional distribution $P(X_i|pa(X_i) = \pi_k)$ is selected from the local credal set $K_P(X_i|pa(X_i) = \pi_k) = [p_{low}, p_{high}]$. One possible cutset is formed by the arcs (1,6), (2,6), (2,8), (3,8), (3,10), (4,10) and (4,12). Removing these arcs we get a polytree associated to a credal set that differs from the original network on local credal sets of variables X_6, X_8, X_{10} and X_{12}. This means that local credal sets $K_P(X_6|X_1, X_2)$, $K_P(X_8|X_2, X_3)$,$K_P(X_{10}|X_3, X_4)$ and $K_P(X_{12}|X_4)$ are approximated by $K_Q(X_j)$, $j = \{6, 8, 10, 12\}$. Assuming that there are no evidence nodes, the expressions for updating Equations (9) and (10) are:

$$\begin{aligned}
\ln Q^*(X_6) &\propto \sum_{X_1,X_2} P(X_1)P(X_2)\ln P(X_6|X_1,X_2),\\
\ln Q^*(X_8) &\propto \sum_{X_2,X_3} P(X_2)P(X_3)\ln P(X_8|X_2,X_3),\\
\ln Q^*(X_{10}) &\propto \sum_{X_3,X_4} P(X_3)P(X_4)\ln P(X_{10}|X_3,X_4),\\
\ln Q^*(X_{12}) &\propto \sum_{X_4} P(X_4)\ln P(X_{12}|X_4).
\end{aligned}$$

From these equations we get approximate credal sets $K_Q(X_6 = 0) = [0.099, 0.346]$, $K_Q(X_8 = 0) = [0.203, 0.664]$, $K_Q(X_{10} = 0) = [0.278, 0.753]$ and $K_Q(X_{12} = 0) = [0.532, 0.810]$. Running the 2U algorithm in this polytree and computing the lower/upper probabilities for all variables, we get a mean square error (MSE), between exact probabilities and those obtained by SV2U, of 2%. Results can be seen in Figure 8. The advantage of the SV2U algorithm over a method like the L2U algorithm [9] is that variational approach gives convergence guarantees that are not offered by L2U.

7 Summary and Conclusions

In this paper we have introduced an explicitly variational approach to inference in credal networks — we have focused on mean field methods and their application to strong extensions. We have investigated the characteristics of such methods and some necessary approximations to make them rely solely on local computations. The paper is thus an initial step in the construction of general approximation methods for credal networks. The next step in this investigation is a more comprehensive study of the empirical characteristics of the SV2U algorithm.

Given the generality and flexibility of variational methods, the set-based methods proposed in this paper seem to be a promising approach to credal networks. Models that contain continuous variables and local credal sets defined by infinitely many constraints are not handled by most existing algorithms; they can in principle be dealt with using variational principles. Hopefully this paper will serve as an initial step in a fruitful avenue of research.

Acknowledgements

The first author has been supported by FAPESP (through grant 02/0898-2). The second author has been partially supported by CNPq (through grant 3000183/98-4). This work has received generous support from HP Brazil R& D.

References

[1] A. Cano and S. Moral. Using probability trees to compute marginals with imprecise probabilities. *International Journal of Approximate Reasoning*, 29:1–46, 2002.

[2] F. G. Cozman. Credal networks. *Artificial Intelligence*, 120(2):199–233, 2000.

[3] F. G. Cozman. Separation properties of sets of probability measures. In *Conference on Uncertainty in Artificial Intelligence*, pages 107–114, 2000.

[4] F. G. Cozman. Graphical models for imprecise probabilities. *International Journal of Approximate Reasoning*, 39(2-3):167-184, 2005.

[5] J. C. da Rocha and F. G. Cozman. Inference with separately specified sets of probabilities in credal networks. In *Conference on Uncertainty in Artificial Intelligence*, pages 430–437, San Francisco, CA, 2002. Morgan Kaufmann.

[6] J. C. da Rocha, F. G. Cozman, and C. P. de Campos. Inference in polytrees with sets of probabilities. In *Conference on Uncertainty in Artificial Intelligence*, pages 217–224, San Francisco, CA, 2003. Morgan Kaufmann.

[7] C. P. de Campos and F. G. Cozman. Inference in credal networks using multilinear programming. In *Proc. of the Starting Artificial Intelligence Researchers Symposium, STAIRS 2004, Valencia - Spain*, pages 50–61, Amsterdam - The Netherlands, 2004. IOS Press.

[8] E. Fagiuoli and M. Zaffalon. 2u: An exact interval propagation algorithms for polytrees with binary variables. *Artificial Intelligence*, 106:77–107, 1998.

[9] J. S. Ide and F. G. Cozman. IPE and L2U: Approximate algorithms for credal networks. In *Proc. of the Starting Artificial Intelligence Researchers Symposium, STAIRS 2004, Valencia - Spain*, pages 118–127, Amsterdam - The Netherlands, 2004. IOS Press.

[10] T. S. Jaakkola. Tutorial on variational approximation methods. *Advanced mean field methods: theory and practice*, 2000.

[11] I. Levi. *The Enterprise of Knowledge.* MIT Press, Cambridge, Massachusetts, 1980.

[12] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 467–475, 1999.

[13] G. Parisi. *Statistical Field Theory.* Addison-Wesley, MA, 1988.

[14] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan-Kaufman, 1988.

[15] L. K. Saul, T. S. Jaakkola, and M. I. Jordan. Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4:61–76, 1996.

[16] L. K. Saul and M. I. Jordan. Exploiting tractable substructures in intractable networks. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 486–492. MIT Press, Cambridge, MA, 1996.

[17] B. Tessem. Interval probability propagation. *International Journal of Approximate Reasoning*, 7:95–120, 1992.

[18] P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London, 1991.

[19] J. Winn. *Variational Message Passing and its Applications.* PhD thesis, Department of Physics, University of Cambridge, Cambridge,UK, 2003.

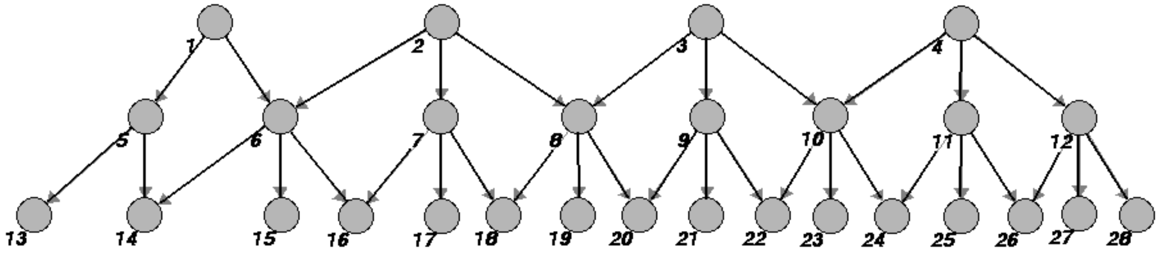


Figure 7: Pyramid network, multilayered graph associated with binary variables, used to test the SV2U algorithm.

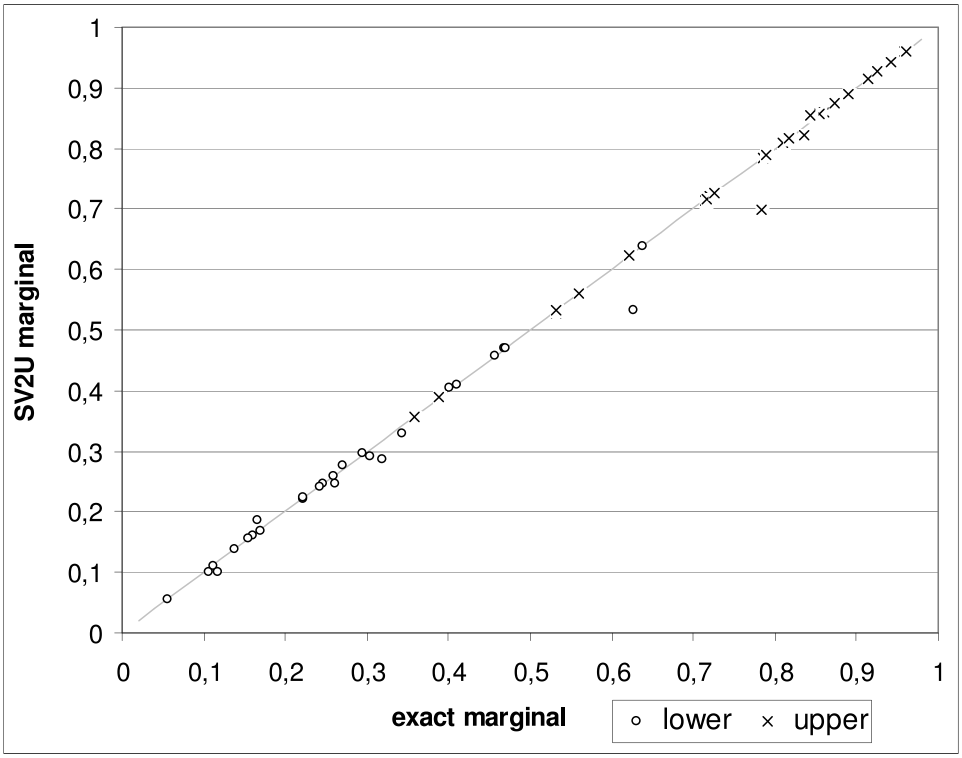


Figure 8: Results of the SV2U algorithm in the Pyramid network. Lower/upper probabilities are computed exactly and approximately, and an MSE=2% is obtained.