FEW 2007

Probabilistic Logics and Probabilistic Networks

Rolf Haenni
Jan-Willem Romeijn
Gregory Wheeler
Jon Williamson

Contents

Part I
Progic

- **A research team carrying out a two-year project.**
- **Aiming to marry logic and probabilistic inferential systems.**
- **In order to bundle the forces of these respective systems.**

1 The Potential of Probabilistic Logic

The Progic project aims to formulate a general logical framework for probabilistic inference, analogous to classical logic.

Classical logic: $a_1, a_2, \ldots, a_n \models b?$

Probabilistic logic: $a_1^{x_1}, a_2^{x_2}, \ldots, a_n^{x_n} \models b^?$

The classical inference concerns truth value assignments, the probabilistic inference concerns probability assignments.

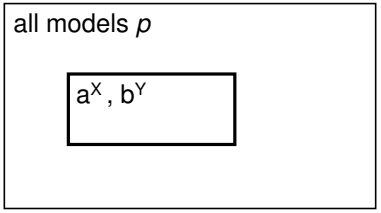
Scope

- Progic covers inductive logic, classical and Bayesian statistics, evidential probability, objective Bayesianism, and probabilistic argumentation theory.
- It has applications in a wide variety of areas: formal epistemology, mathematical statistics, the philosophy of science, artificial intelligence, bioinformatics, linguistics, psychometrics.

Progic strategy

First we show how a number of systems for probabilistic inference can be unified in the Progic framework.

- The key question of each is representable as $a_1^{X_1}, a_2^{X_2}, \ldots, a_n^{X_n} \models b^?$.
- Each provides semantics, a notion of model, for this general question.

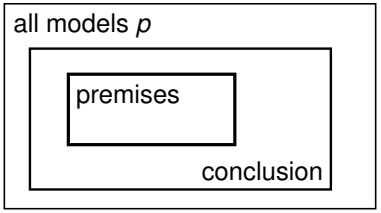


Then we show how probabilistic nets can be used to reduce the computational complexity of the inferences that the systems have in common.

- Convex sets of probability assignments are typically all that is needed.
- Credal and Bayesian nets can handle these very well.

2 Standard Probabilistic Semantics

Models are characterised by probability assignments. Premises and conclusions are constraints on models. An inference is valid if satisfaction of the conclusion constraint is guaranteed by the combined constraints of the premises.



This notion of validity is the common core to all probabilistic logics.

Probability theory as logic

A probability space is a tuple, $(W, \mathcal{F}, p)$, where $\mathcal{F}$ is a σ-algebra over a set W and $p : \mathcal{F} \longrightarrow [0, 1]$ is a probability measure defined on the algebra $\mathcal{F}$ satisfying

P1. $p(\emptyset) = 0$, $p(W) = 1$

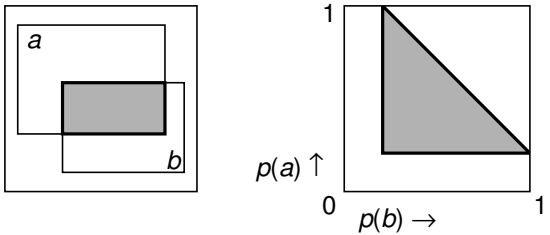
P2. $p(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=i}^{\infty}(A_i)$, when A_i are countable, pairwise disjoint elements of $\mathcal{F}$.

The probability space provides the models. They relate to a langauge as follows.

- We associate each $w \in W$ with a truth assignment on atomic propositions a_i in a language L by the indicator function $I(w, a) \in \{0, 1\}$.
- We can identify the set $A_i^1 = \{w : I(w, a_i) = 1\}$ with the proposition a_i, and similarly $A_i^1 \wedge B^1 = \{w : I(w, a_i) \times I(w, b) = 1\}$.
- The expressions a_i^q and $p(A_i^1) = q$ refer to the probability assignment in the syntax and the semantics respectively.

Sets of probabilities

A valid inference in the standard semantics is $(a \wedge b)^{0.3}, (\neg a \wedge b)^{0} \models b^{0.3}$. A slightly more involved example is $(a \wedge b)^{0.3} \models b^{Y?}$. Now the premises do not constrain the conclusion to a sharp value of $p(b) = Y$ but rather to a convex set of probability values.



In the standard semantics, both premises and conclusion may constrain the probability assignments only up to a convex set of probabilities: $X, Y \in [\underline{p}, \overline{p}]$. The inference problem is to find the smallest upper and the largest lower bound to Y given the intervals X.

3 Credal and Bayesian Networks

We can represent much more elaborate inferential systems in the Progic framework, and interpret the inference problem of the framework in terms of these inferential systems.

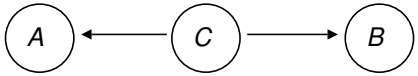
- The different inferential systems provide different ways for determining the set of models that comply to the premises.
- These representations and interpretations allow us to unify different inferential systems, and bring out their common core.

Often probabilistic inference systems are computationally intractable. However. . .

- We provide an efficient inferential procedure for the main problem in the Progic framework using so-called credal networks.
- We thereby serve all the inferential systems that can be accommodated in the framework.

Bayesian networks

A Bayesian network is a representation of a probability function over random variables that captures the independence relations among these variables graphically.



$$p(C^1) = 0.3$$

$$p(A^1|C^1) = 0.7, \quad p(A^1|C^0) = 0.9$$

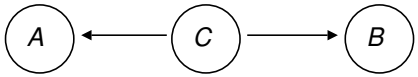
$$p(B^1|C^1) = 0.1, \quad p(B^1|C^0) = 0.2$$

The independencies are laid down in the Markov Condition: $U \perp\!\!\!\perp ND_U \mid Par_U$. For the above chain we can write

$$p(A, B, C) = \prod_{U \in \{A,B,C\}} p(U \mid par_U).$$

Credal networks

A credal net represents a so-called credal set: a closed convex set of probability functions.



$$p(C^1) \in [0.3, 0.32]$$

$$p(A^1|C^1) \in [0.7, 1], \quad p(A^1|C^0) \in [0.1, 0.9]$$

$$p(B^1|C^1) = 0.1, \quad p(B^1|C^0) \in [0.2, 0.5]$$

A credal network determines a set of credal sets. The specific extension of the network determines the independence assumptions that the members of the credal set satisfy.

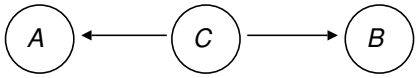
Natural: include every probability for which the conditional probabilities in the network are within the bounds, so no independence assumptions.

Strong: assume the independence for the extremal points, and then take the convex hull.

Complete: take all the Bayesian nets for which the conditional probabilities lie within the bounds, so complete independence.

Parameterised credal nets

A parameterised credal net represents a credal set in which the extremal points are interrelated. The relations arise when constraints involve more than one node in the network, for example A and C.



$$\gamma \stackrel{\text{df}}{=} p(C^1) \in [0.3, 1]$$

$$p(A^1|C^1) = \frac{0.3}{\gamma}, \quad p(A^1|C^0) = 0$$

$$p(B^1|C^1) = 0.1, \quad p(B^1|C^0) \in [0.2, 0.5]$$

Parameterised credal nets offer the same advantages, and allow for the same computational procedures as ordinary credal nets. But there are some restrictions to the possible functional relations between interval bounds.

Progic inferential procedure
The inference problem in the Progic framework is to find the minimal Y such that

$$a_1^{X_1}, a_2^{X_2}, \ldots, a_n^{X_n} \models b^Y.$$

Credal networks can be used to speed up this process. The strategy is as follows.

Step 1: Employ the specifics of the inferential systems that can be represented in the Progic framework to determine a probabilistic net. This will vary according to the inferential system that determines the semantics.

Step 2: Use the network from Step 1 to calculate Y efficiently. This step is independent of the chosen semantics. It uses a hill-climbing algorithm on the contours of the credal set to find the upper and lower bounds for Y.

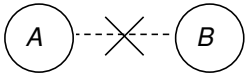
Hill climbing

Some background to the hill-climbing algorithm used for the second step.

1. Transform b into an equivalent *disjoint* DNF $b_1 \vee \cdots \vee b_s$, for which $b_i \wedge b_j \equiv \bot$ if $i \neq j$.
   - For example, $b = a_1 \vee a_2$ is transformed into $b_1 = a_1$ and $b_2 = \neg a_1 \wedge a_2$.
   - Note that $p(B^1) = p(B^1_1) + \cdots + p(B^1_s)$.
2. Perform inference in the credal network.
   - Calculate lower and upper bounds $\underline{p}(B^1)$ and $\overline{p}(B^1)$.
   - Very inefficient in general, so approximation is indispensable.
   - Logical compilation: expensive offline phase, cheap online phase.
     - Compile the credal net using techniques from Bayesian nets.
     - Instantiate the compiled net for all disjoint queries b_j.
     - Apply hill-climbing to minimize and maximize $p(B^1) = \sum_{j=1}^{s} p(B^1_j)$.
3. Use $\underline{p}(B^1)$ and $\overline{p}(B^1)$ as bounds for Y in b^Y.

4 Networks for the Standard Semantics

The natural extension of a credal net comprises of all probability functions over $\{A, B\}$ for which the restrictions on conditional probabilities hold.



In this case, A and B are independent. Imagine further that we have the following premises:

$$a^{[0.25,075]}, \quad b^{[0.50,1]}.$$

In terms of the probability assignments in the semantics:

$$p(A^1) \in [0.25, 0.75], \quad p(B^1) = [0.50, 1].$$

Under the natural and strong extension, the resulting credal set includes probability functions for which the conditional independence suggested by the network does not hold.

Strong versus complete extension

It is only when we assume what may be called the complete extension of a credal net that these independencies hold. Assuming the complete extension means adding the following premise to the scheme:

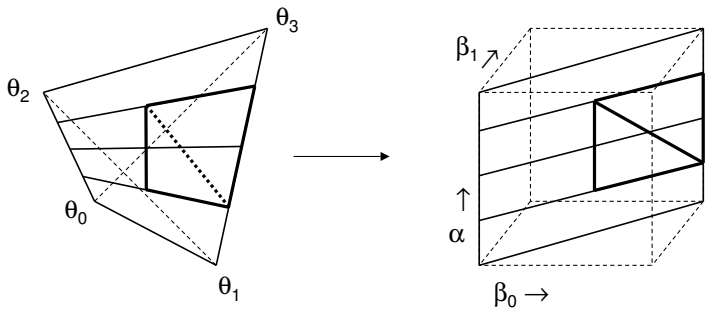
$$\forall \alpha, \beta \in [0,1] : a^{\alpha} \wedge b^{\beta} \wedge (a \wedge b)^{\alpha\beta}.$$

This comes down to the following restriction to the set of probability assignments:

$$p(A^1) = \frac{p(A^1 \cap B^1)}{p(B^1)}.$$

But we can already run efficient algorithms with credal networks on the assumption of the strong extension: we employ independence to cover cases for which independence does not hold.

Interestingly, it depends on the parameterisation, or the metric, of the space of probability assignments whether assuming the strong extension is the same as assuming the complete extension or not.

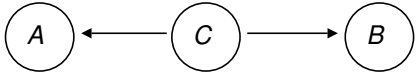


Using credal nets

In order to employ the computational advantages of credal networks, we must minimally assume the strong extension. But nothing in the standard semantics itself warrants any independence assumptions.

Dilation

It is notable that the independence relations are not only of computational use. They can do conceptual work in the dynamics of probability intervals, in particular to avoid a phenomenon called dilation. We do this by confining p to the complete extension of the credal network.



$$p(C^1) \in [0, 0.5], \quad p(A^1|C^1) = 0.5$$
$$p(B^1|C^0) \in [0.5, 1], \quad p(B^1|C^1) \in [0, 0.5]$$

Example: say that we learn c and condition on it, so that our belief in a is $p(A^1|C^1)$, and that we are then offered a test on b. After learning either that b or that $\neg b$, the probability assignment to a is again the whole interval, because $p(A^1|C^1 \wedge B^j) \in [0, 1]$ for $j = 0, 1$. So by learning whether b we invariably loose all information on a.

Part II

Statistical Inference and Evidence

- **The Progic framework can be applied to statistical inference.**
- **with the aim of providing statisticians with additional logical tools.**
- **and widen the view on evidential relations.**

5 Statistical Inference

An important application of probability theory is the use of statistics in science, predominantly classical statistics as devised by Fisher and Neyman and Pearson.

Classical statistics
Classical statistical procedures concern probability assignments $p_H(E)$ over samples E relative to a statistical hypothesis H.

- Neyman-Pearson test function: $\frac{p_{H_0}(E)}{p_{H_1}(E)}$.
- Fisher estimate: parameterise the hypotheses with θ, then the estimate is

$$\{\theta : \forall\theta'(p_{H_{\theta'}}(E) \leq p_{H_\theta}(E))\}.$$

Can we faithfully accommodate classical statistics in the inferential schema of Progic? Perhaps classical statistical procedures cannot be seen as inferences to start with. Rather they are a guide for making decisions, which have certain error rates associated with them.

The fiducial argument

Fisher suggested a way of capturing classical statistics in terms of a probabilistic inference by means of so-called fiducial probability.

Dawid and Stone provide a general characterisation of the set of statistical problems $\langle H_\theta, E \rangle$ that allow for application of the fiducial argument, using so-called functional models:

$$f(\theta, \omega) = E$$

$$V_E(\theta) = \{\omega : f(\theta, \omega) = E\}$$

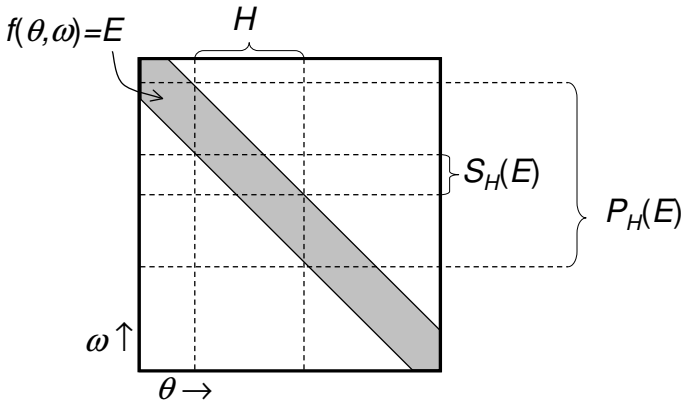
$$p(E|H_\theta) = \sum_{\omega \in V_E(\theta)} p(\omega)$$

A specific class of functional models allows for the application of the fiducial argument:

$$\forall \theta \neq \theta' : \; V_E(\theta) \cap V_E(\theta') = \emptyset \quad \Rightarrow \quad p(\theta) = \sum_{\omega \in V_\theta(E)} p(\omega).$$

Support and possibility

Functional models show the limits of the fiducial argument, but as Kohlas and Monney (200X) show, they also provide the starting point for an adapted and more general version of the fiducial argument. The general idea is best captured in a picture.



Formal explication

Based on the functional model $f(H_\theta, \omega) = E$, a hypothesis $H = \cup_{\theta \in I} H_\theta$, determined by an interval I, can always be assigned a degree of support (Sup) and possibility (Pos). With $U_E(\omega) = \{H_\theta : f(\theta, \omega) = E\}$, we define

$$S_H(E) = \{\omega : U_E(\omega) \subset H\} \qquad Sup(H) = \sum_{\omega \in S_H(E)} p(\omega),$$

$$P_H(E) = \{\omega : U_E(\omega) \cap H \neq \emptyset\} \qquad Pos(H) = \sum_{\omega \in P_H(E)} p(\omega).$$

This inference can be captured in the inferential scheme of Progic as follows:

$$(f : f(\theta, \omega) = E\})^1 \wedge \omega^{p(\omega)} \wedge e^1 \models H^{[Sup(H), Pos(H)]}.$$

6 Networks for Statistical Inference

Credal networks apply to this version of classical statistical inference in a number of ways. The basic idea is that we can exploit independence relations inherent to the set-up of functional models.

Credal networks

- as tools to structure functional models;
- as used in the standard semantics;
- to express vague evidence.

Independence in functional models

One way to employ networks is by identifying and exploiting the independence relations between statistical parameters that appear in the functional models:

$$f(\theta_1, \theta_2, \omega) = g_1(\theta_1, \omega) g_2(\theta_2, \omega).$$

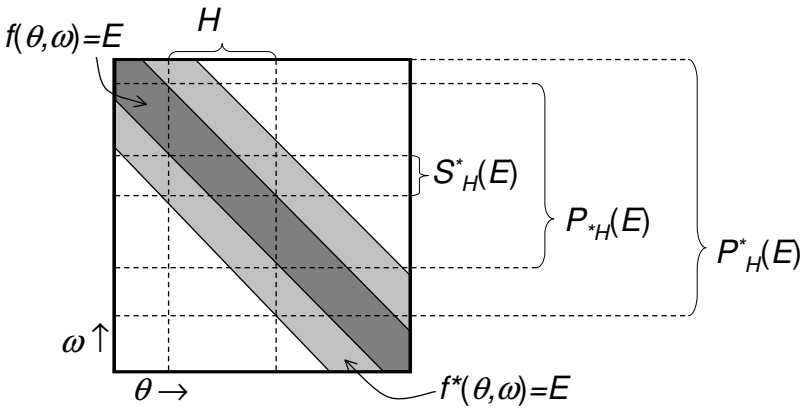
Logical combinations of hypotheses
The conclusion $H^{[Sup(H),Pos(H)]}$ can be employed in derivations of the probability of logical combinations of several hypotheses. The Progic framework here provides inference procedures based on credal networks.

Yet we must be careful in taking the intervals of *Sup* and *Pos* as interval-valued probability assignments simpliciter.

- The functions *Sup* and *Pos* express the probability of nested but different events.
- Combining hypotheses on the level of the underlying functional models may lead to different results than combining the hypotheses with interval-valued probability directly.
- This signals that not all of the semantics of classical statistics is covered by the standard semantics inherent in credal networks.

Vague evidence

An interesting possibility is to introduce vagueness into the evidence, $V_E(\theta) \subsetneq V^\star_E(\theta)$. This leads to intervals for both support and possibility separately. The separate interval-valued degrees of support and possibility do not suffer from the above defects.



7 Bayesian Statistical Inference

Bayesian statistics is much more easily connected to the inferential schema of Progic.

Second-order probability
The distinguishing feature of Bayesian statistical inference is that it assigns probability over statistical hypotheses. The inferences are captured in

$$\forall j \leq n : h_j^{p(H_j)} \wedge (e|h_j)^{\theta_j} \models (h_j|e)^{p(H_j|E)}.$$

The schema combines probabilistic premises, namely the priors and likelihoods of hypotheses, to arrive at probabilistic conclusions, namely a conditional posterior over the hypotheses.

Some notation

Arguments of the form $(a|b)^\gamma$ are not normal expressions in the language. But at bottom they are restrictions to a set of probability assignments, or models for short.

$$(a|b)^\gamma \iff \forall \beta \in [0,1] : p(B^1) = \beta, \quad p(A^1 \cap B^1) = \beta\gamma.$$

In terms of the premises in the language:

$$(a|b)^\gamma \iff \forall \beta \in [0,1] : b^\beta \wedge (a \wedge b)^{\beta\gamma}.$$

With this interpretation, the Bayesian inference follows directly from the standard semantics.

A continuum of hypotheses

Many statistical applications employ a continuum of hypotheses H_θ. The inference then involves an uncountable infinity of premises and conclusions. But by choosing $\theta_j = \frac{2j-1}{2n}$ we can approximate the continuous model arbitrarily close by increasing n.

Exchangeability and inductive logic

Predictions $p(E'|E)$ can be derived from the posterior probability assignments $p(H_\theta|E)$ and the likelihoods for E'. Such predictions also fit the framework:

$$\forall\theta : H_\theta^{p(\theta)} \wedge (E|H_\theta)^{\theta_E} \wedge (E'|H_\theta)^{\theta_{E'}} \models (E'|E)^{p(E'|E)},$$

Using exchangeability, we can represent any such statistical inference on the basis of multinomial distributions in terms of a finite number of probability assignments over observations:

$$\forall\pi : \pi(E \cap E')^{p(E\cap E')} \models (E'|E)^{p(E'|E)}.$$

Here π is an order permutation. Carnap, Jeffrey, and others consider special cases of the finite reformulation in what has become known as inductive logic.

Interval-valued priors

Intervals of probability assignments constitute a wider set of restrictions to the probability assignments. They can be employed in Bayesian statistical inference in at least two ways.

- Walley shows that we can allow for interval-valued assignments to statistical hypotheses. They can be dealt with adequately by considering a range of prior density functions over the hypotheses. Any range of priors leads to so-called hyper-Carnapian prediction rules.
- From the detailed knowledge of a sharp-valued probability assignment over hypotheses we may derive interval-valued probability assignments for θ. Example:
  - fix u and l such that $\int_0^l p(H_\theta|E)d\theta = \int_u^1 p(\theta|E)d\theta = 0.025$;
  - the inferential scheme may then take on the form

$$\theta^{[.01,.99]} \wedge (E|H_\theta)^{\theta_E} \models (H_\theta|E)^{[.08,.13]};$$

This inferential form is elliptic: the premise $\theta^{[.01,.99]}$ does not fix the detailed shape of the prior probability $p(H_\theta)d\theta$, but we need this detailed form to arrive at the specific conditional credence interval $H_\theta^{[.08,.13]}$.

8 Networks for Bayesian Statistical Inference

There are again various ways in which credal networks may be employed in expanding and improving standard Bayesian statistical inference.

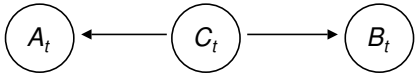
Credal networks

- as tools to structure a statistical model;
- as used in the standard semantics;
- to express uncertain evidential relations.

Models as credal networks

A credal set can be viewed a statistical model: each member is a probability function over some set of variables. Any credal set may be captured by a second-order probability over all probability functions over the variables that is non-zero only at functions belonging to the credal set.

The representation of the model as a credal network is useful when adapting the probability over the model in learning from data. Consider subsequent observations, at times t, of three binary variables $U_t = \{A_t, B_t, C_t\}$.



A statistical hypothesis on these variables must fix $2^3 - 1 = 7$ free probabilities. But the complete extension of the credal net may be parameterised by a 5-tuple $\eta = \langle \gamma, \alpha_0, \beta_0, \alpha_1, \beta_1 \rangle$, where

$$p(C_t^1 | H_\eta) = \gamma \quad \text{with} \quad \gamma \in [0, 1],$$

$$p(A_t^1 | C_t^i \wedge H_\eta) = \alpha_i \quad \text{with} \quad \alpha_i \in [0, 1],$$

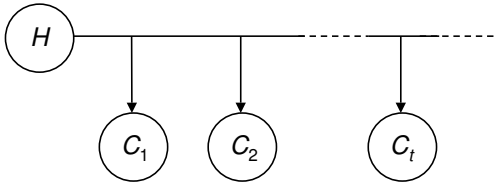
$$p(B_t^1 | C_t^i \wedge H_\eta) = \beta_i \quad \text{with} \quad \beta_i \in [0, 1].$$

This reduction in the dimensions of the model entails major reductions in computational load.

Logically complex statistical hypotheses

We may include statistical hypotheses as hidden nodes in a credal network. This allows us to derive logical combinations of statistical hypotheses by the inference machinery of Progic.

Example: a node H_γ with values $\gamma \in [0, 1]$ may be added as a common parent to instantiations, at specific t, of the single binary variable C_t.



But we must be careful in interpreting the interval-valued probability assignments to statistical hypotheses that result from such inferences.

Uncertain evidential relations

Consider the hypotheses H_j for $j = 0, 1$, include the hypothesis node H_j in the credal network, and replace the sharp probability values for $C_t^1 = 1$ with

$$\begin{aligned} p(C_t^1|H_0) &\in [0.3, 0.7], \\ p(C_t^1|H_1) &\in [0.6, 0.8]. \end{aligned}$$

The common machinery of credal networks can be applied directly to such interval-valued likelihoods.

The interpretation of this is that the statistical hypotheses are not exactly clear on the probability of C_t^1, although they do differ on it. This formal possibility is in a sense complementary to the well-known method of Jeffrey conditioning.

9 Conclusion

Part I : We may represent and interpret many different probabilistic inferential systems in the Progic framework.

Part II : Efficient inference becomes possible with the use of credal networks. The required independence assumptions are motivated by the different systems.

Acknowledgements
We are very grateful to The Leverhulme Trust for supporting Progicnet. See:

www.kent.ac.uk/secl/philosophy/jw/2006/progicnet.htm