JON WILLIAMSON

PROBABILITY LOGIC

Practical reasoning requires decision-making in the face of uncertainty. Xenelda has just left to go to work when she hears a burglar alarm. She doesn't know whether it is hers but remembers that she left a window slightly open. Should she be worried? Her house may not be being burgled, since the wind or a power cut may have set the burglar alarm off, and even if it isn't her alarm sounding she might conceivably be being burgled. Thus Xenelda can not be certain that her house is being burgled, and the decision that she takes must be based on her degree of certainty, together with the possible outcomes of that decision.

If Xenelda, or X for short, uses classical logic to make a decision, she will not get very far. Classical logic has no explicit mechanism for representing the degree of certainty of premises in an argument, nor the degree of certainty in a conclusion, given those premises. X must look for a logic that can represent uncertainty.

Here we will look to probability as a representation of uncertainty, and see how a logic of practical reasoning might involve probability, in order to help agents like X. The plan is this: first we shall recall the standard definition of probability, and the standard interpretations of this formal definition. Next we shall look at attempts to incorporate probability into a logic, and go on to investigate a practical question, namely inference.

1 PROBABILITY AND ITS INTERPRETATIONS

Probability has a rather technical mathematical characterisation. Given an arbitrary non-empty space Ω and a σ-field $\mathcal{F}$ of subsets of Ω (that is, a nonempty class of subsets of Ω closed under complement and countable unions), $p : \mathcal{F} \longrightarrow \mathbf{R}$ is a *probability measure* if (for all $a, b, a_1, a_2, \ldots \in \mathcal{F}$)

M1: $0 \leq p(a) \leq 1$;

M2: $p(\emptyset) = 0$ and $p(\Omega) = 1$;

M3: if $a_1, a_2, \ldots$ are disjoint then $p(\bigcup_{i=1}^{\infty} a_i) = \sum_{i=1}^{\infty} p(a_i)$.

A new function, *conditional probability* is induced by probability according to the following definition:

MC: if $p(a) > 0$ then $p(b|a) = p(a \cap b)/p(a)$.

For example, let B stand for burglary and B' for no burglary. Suppose $\Omega = \{B, B'\}$, and $\mathcal{F} = \{\emptyset, \{B\}, \{B'\}, \{B, B'\}\}$. Then the axioms require that $p(\{B\}) + p(\{B'\}) = 1$ and the definition of conditional probability ensures that $p(\{B\}|\{B'\}) = 0$.

This gives a formal definition of probability,[1] but it doesn't tell us what probability means. Probability has a number of standard interpretations, and we shall take a brief look at these now.[2] As a starting point the elements of the domain of a probability measure are usually called events. Just what the events are depends somewhat on the interpretation of probability.

1.1 Subjective degrees of belief

Ramsey[3] and de Finetti[4] interpreted probability as degree of rational belief. According to this interpretation an agent, X for example, has a belief function p_X^τ at time τ over the domain of events, which are single-case (that is, unrepeatable). X's house being burgled that morning would qualify as such an event. Thus for $a \in \mathcal{F}, p_X^\tau(a)$ measures X's degree of belief at τ in the occurrence of event a. In our example, $p_X^\tau(\{B\})$ represents X's degree of belief at time τ that a burglary has taken place that morning. One can determine X's degrees of belief by analysing how she is prepared to bet:

- $p_X^\tau(a) = x$ if and only if at time τ, X is willing to bet $x\Delta\Theta$ on event a occurring, with return $\Delta\Theta$ if a does occur, where $\Delta\Theta$ is an unknown stake (either monetary or in terms of some measure of utility) which may depend on $p_X^\tau(a), \Theta \in \mathbf{R}_{\geq 0}$ being the magnitude and $\Delta = \pm 1$ the direction of the stake; and

- $p_X^\tau(b|a) = x$ iff at τ, she is prepared to bet $x\Delta\Theta$ on b occurring, with return $\Delta\Theta$ if both a and b occur but with the bet being called off if a fails to occur.

X's belief function p_X^τ is deemed to be *coherent* if no stake-maker can choose stakes which make X lose money whatever happens. By the Dutch book argument,[5] p_X^τ is coherent if and only if it is a probability measure. Thus probability measures are coherent belief functions. This gives a subjective interpretation of probability in the sense that probabilities are associated with a subject's state of knowledge or belief, as opposed to attaching directly to the physical world as stipulated by objective interpretations.

The notion of coherence can be used to argue for further rational constraints on X's belief function. For example one can apply a diachronic

[1] [Kolmogorov, 1933] was a key pioneer behind the mathematical theory of probability — see [von Plato, 1994] for further historical details and [Billingsley, 1979] for a good exposition of the modern theory.

[2] See [Howson, 1995] for a more detailed survey.

[3] [Ramsey, 1926].

[4] [de Finetti, 1937].

[5] [Ramsey, 1926] and [de Finetti, 1937]. See also [Williamson, 1999] regarding the axiom of countable additivity [*M3*].

Dutch book argument to show that if $p_X^\tau(a) > 0$ and between times τ and $\tau + 1$ X comes to learn only of the occurrence of event a, then her belief function should change via *Bayesian conditionalisation*:

- $p_X^{\tau+1}(b) = p_X^\tau(b|a)$.

Thus one can bolster the subjective theory by adding further rationality requirements — although proponents of the subjective interpretation often disagree as to which extra principles should be adopted.[6]

Another point of disagreement boils down to attitudes toward other interpretations of probability. *Strict subjectivists* like de Finetti argue that subjectivism is the only viable interpretation of probability, while others are tolerant of, or argue in favour of, one or more of the following objective interpretations.

1.2 Objective frequencies

Von Mises was the first to work through the idea that probabilities are measures of frequency.[7] According to this interpretation an event a in $\mathcal{F}$ can be repeatably instantiated, *a burglary* rather than *X's burglary that morning*, and the *frequency* of a may be defined as its limiting relative frequency in a *collective*. This collective is a denumerable sequence of mutually exclusive and exhaustive elements of $\mathcal{F}$. In our example a collective may look like $(\{B'\}, \{B\}, \{B'\}, \{B\}, \{B'\}, \{B'\}, \{B'\}, \ldots)$ and may be obtained from examining X's house each time its alarm sounds to see if it was burgled. Von Mises invoked two empirical laws, the first of which claimed that for any naturally occurring collective, the relative frequency of an event a in the first n places of the collective tends to a limit, the frequency of a, as n increases. Thus in the above collective the relative frequency of $\{B\}$ progresses $0, \frac{1}{2}, \frac{1}{3}, \frac{1}{2}, \frac{2}{5}, \frac{1}{3}, \frac{2}{7}, \ldots$ and is assumed to converge to a fixed limit. The second empirical law claims that this frequency is constant over all subsequences selected by recursive place selections from the collective — so for example if we form a new collective by taking all the even places of the original collective, the limiting frequency in the new collective is the same as that in the original collective.[8] From these empirical laws von Mises' deduced that frequency obeys the axioms of *finitely additive* probability, that is, the above axioms with [*M3*] replaced by finite additivity, which is the special case in which finitely many attributes are considered.[9]‘

Popper noted that there are intuitive difficulties concerned with the collective of outcomes of rolls of a biased die interspersed with one or two rolls

[6] Those who advocate Bayesian conditionalisation usually call themselves ‘Bayesians’.

[7] [von Mises, 1928].

[8] See [von Mises, 1964] for the details.

[9] Note however that von Mises later included countable additivity as an extra stipulation.

of a fair die. While it makes sense to say that the frequency of a 5 is $\frac{1}{4}$ for such a collective, intuitively the frequency changes according to which die is rolled.[10] Thus von Mises' theory is often amended to a *propensity* theory.[11] This states that it is reasonable only to consider collectives generated by a repeatable experiment, such as rolling a particular die, and that once this move is made, the frequency of an event is dependent only on the generating experiment. In other words the frequency is constant over all collectives generated by the same repeatable experiment. This might be given the status of an empirical law, or, as with Popper, such a concept of probability may be taken as a scientific primitive.

Another version of the frequency theory takes issue with the law that says the relative frequency in a collective converges to a fixed limit. One can argue that one should only assume that relative frequency converges *in probability* to the limit, which means that for a small proportion of collectives (a set of measure 0) there will be no convergence.[12] Another frequency theory, *actual frequency*, takes issue with the demand that collectives be infinite, which usually fails for outcomes of real experiments, and defines frequency to be the relative frequency in the actual (often finite) sequence of outcomes.[13]

1.3 Objective chances

We obtain quite a different objective notion of probability if we demand that events in the domain of a probability measure are single-case (as they were in the subjective theory) as opposed to the repeatable events of the frequency theory. We can merely hypothesise that the mathematical theory has such an objective physical interpretation, and find ways to test this assertion in order to confirm or refute it. Thus while many versions of the last two interpretations offer an explicit definition of probability in terms of observable beliefs and frequencies respectively, this theory implicitly defines objective single-case probabilities, or *chances*, and requires some form of prediction method to test the definition.[14] There are two standard ways of drawing predictions from chances. One can claim that chances give rise to certain frequencies as experiments are repeated, and then test to see

[10] [Popper, 1983] pp. 352–356.

[11] See for example [Popper, 1972]. The idea behind propensity theories, like that behind frequency theories, goes back a long way — see [Peirce, 1910] paragraph 664.

[12] The mathematical theory of probability only implies convergence according to the laws of large numbers, and consequently one can only account for the phenomenon of frequency *from within* the mathematical theory if one assumes the weaker notion of convergence in probability. See [Kolmogorov, 1933], [Neapolitan, 1992].

[13] See [Williamson, 1999b].

[14] Note that Popper's formulation of the propensity theory, while not a single-case theory, made use of this type of implicit definition of probability rather than von Mises' operationalist definition involving empirical laws.

whether those frequencies are obtained. Thus if the chance of X's house being burgled this morning, given that her alarm is sounding, is $\frac{1}{3}$, then one can claim that if we form a collective by checking for burglary each time X's alarm sounds, the frequency of burglary will be $\frac{1}{3}$ in this collective. Alternatively one can claim that the chance at time τ of an event is the degree to which an agent should believe at τ that it will occur, were she to have at τ all the relevant information pertaining to the occurrence of the event. We might then test this claim by seeing if the agent can be made to lose money by betting according to chances. Thus in our example X should believe that her house is being burgled, given that her alarm is sounding and other relevant information, to degree $\frac{1}{3}$.[15]

Having sketched the concept of probability and its interpretations, we shall move on to the relationship between probability and logic, which is the central theme of this chapter. In the next section we will look at attempts to integrate probability and logic. Later we shall investigate the important practical problems to do with probabilistic representation and inference.

2 PROBABILITY AS LOGIC

The starting point for most theories that integrate probability and logic is to attach probability to logical statements rather than events. The motivation here is that logic operates on statements and so it would be natural if a probabilistic logic for practical reasoning were to do the same. A second motivation is that when we look at the application of the mathematical theory of probability, we see that probabilities are usually posited of a random variable taking a certain value. Such expressions are more naturally thought of as statements of the form $X = x$ than events of the form $\{\omega \in \Omega : X(\omega) = x\}$.[16]

Unfortunately, confusion is often the upshot of the move from events to sentences. The problem is that the literature contains a plethora of new axiomatisations of probability on sentences, few of which bear a clear resemblance the mathematical formulation of probability.[17] One can however make the link between probability on sentences and the mathematical theory more perspicuous, as follows.

[15] See [Mellor, 1971] and [Lewis, 1980] for detailed defences of the chance approach.

[16] [Scott and Kraus, 1966], p. 219.

[17] Thus it requires significant effort to work out how the new notions of probability relate to the mathematical notion — see [Roeper and Leblanc, 1999] for a glimpse of what is required. One reason for the abundance of axiomatisations is that probability theory is often used to give a new semantics for logic, whereby a statement is logically true if given probability 1 by all probability functions, in which case axiomatisations of probability must be *autonomous*, in that they must not themselves involve logical notions. [Popper, 1934] appendix *iv and [Field, 1977] provide examples of such axiomatisations and this approach. However, our task is to investigate probability as an extension rather than a replacement of logic, so we need not enter into the intricacies of these axiomatisations.

Consider a propositional language $\mathcal{L}$ involving a countable set of propositional variables $\{c_1, c_2, \ldots\}$. Sentences $\mathcal{S}^\infty\mathcal{L}$ are formed by applying the usual connectives $\neg, \vee, \wedge, \rightarrow, \leftrightarrow$ and allowing denumerable disjunctions and conjunctions $\bigvee, \bigwedge$ (such sentences are known as *infinitary*).[18] The usual semantics (including logical truth, logical implication $\models$ and logical equivalence $\equiv$) is available if we consider the (uncountable) infinity of truth functions. Then $p : \mathcal{S}^\infty\mathcal{L} \longrightarrow \mathbf{R}$ is a *probability measure* if (for all $\theta, \phi, \theta_1, \theta_2, \ldots \in \mathcal{S}^\infty\mathcal{L}$)

L1: $0 \leq p(\theta) \leq 1$;

L2: $p(\theta \wedge \neg\theta) = 0$ and $p(\theta \vee \neg\theta) = 1$;

L3: if $\theta_1, \theta_2, \ldots$ are mutually exclusive[19] then $p(\bigvee_{i=1}^\infty \theta_i) = \sum_{i=1}^\infty p(\theta_i)$;

and *conditional probability* is defined by:

LC: if $p(\theta) > 0$ then $p(\phi|\theta) = p(\theta \wedge \phi)/p(\theta)$.

Note that one consequence of the axioms is that if two sentences are logically equivalent, $\theta \equiv \phi$, then they have the same probability.

The clear connection between the mathematical and the logical axioms of probability is due to the formal fact that the *Lindenbaum algebra* $\mathcal{S}^\infty\mathcal{L}/\equiv$ is a σ-algebra which is isomorphic to a σ-field $\mathcal{F}$ of subsets of the space Ω of truth functions. Thus each sentence θ corresponds to an element a of $\mathcal{F}$ and under this mapping the logical axioms on $\mathcal{S}^\infty\mathcal{L}$ are equivalent to the mathematical axioms on $\mathcal{F}$.

This works according to the following construction. The space Ω of truth functions is just the space of binary sequences. A truth function ω is of the form $(t_1(\omega), t_2(\omega), \ldots)$ where $t_i(\omega) \in \{T, F\}, i = 1, 2, \ldots$, signifies the truth value of c_i. A *cylinder of rank n* is of the form $a = \{\omega : (t_1(\omega), \ldots, t_n(\omega)) \in H\}$, where $H \subseteq \{T, F\}^n$. The set of cylinders of all ranks is a field,[20] and we shall consider the σ-field $\mathcal{F}$ generated by this field. Now to each sentence θ in $\mathcal{S}^\infty\mathcal{L}$ there corresponds an element of $\mathcal{F}$ which contains just the truth functions that satisfy θ. This can be shown inductively. If θ is propositional variable c_n, take H as $\{T, F\}^{n-1} \times \{T\}$ and then the corresponding cylinder of rank n contains just the truth functions that satisfy θ. If θ is $\neg\phi$ and $a \in \mathcal{F}$ contains the satisfiers of ϕ, then the complement of a, a', which is also in $\mathcal{F}$, will contain the satisfiers of θ. The union of the satisfiers of θ and ϕ will be the satisfiers of $\theta \vee \phi$, and a countable union is required for $\bigvee_{i=1}^\infty \theta_i$. Other connectives can be reduced to these. The converse is also true: to each $a \in \mathcal{F}$ there is a $\theta \in \mathcal{S}^\infty\mathcal{L}$ which is satisfied by just the truth functions in a. If a is a cylinder with places $i_1, \ldots, i_k$ fixed to T and places $j_1, \ldots, j_l$

[18] See [Karp, 1964].

[19] In the sense that $\models \neg(\theta_i \wedge \theta_j)$ for all i and j.

[20] [Billingsley, 1979], p. 27.

fixed to F, then $c_{i_1} \wedge \ldots \wedge c_{i_k} \wedge \neg c_{j_1} \wedge \ldots \wedge \neg c_{j_l}$ is the required sentence, and a countable union of cylinders requires a countable disjunction of such sentences. Now θ and ϕ both correspond to the same $a \in \mathcal{F}$ iff they are satisfied by the same set of truth functions, i.e. $\theta \equiv \phi$. Thus we have a bijection between $\mathcal{S}^\infty\mathcal{L}/\equiv$ and $\mathcal{F}$.

Such an infinitary propositional system is a useful formalism because it clarifies the link between probability on sentences and probability on events. However, many are cautious about using infinitary systems, either because they find the concept of infinitary sentences counterintuitive or for technical reasons — for example only a weak version of the completeness theorem can be proved. Interestingly, if we restrict attention to a finitary language and define probability accordingly then no probabilistic information is lost, in the sense that there are no more probability measures on an infinitary language than on the finitary language which it extends. Let $\mathcal{SL}$ denote the finitary sentences (involving the same countable set of propositional variables $\mathcal{L}$, but no countable disjunctions or conjunctions). Let $p : \mathcal{SL} \longrightarrow \mathbf{R}$ be a *(finitary) probability measure* if (for all $\theta, \phi \in \mathcal{SL}$)

F1: $0 \leq p(\theta) \leq 1$;

F2: $p(\theta \wedge \neg\theta) = 0$ and $p(\theta \vee \neg\theta) = 1$;

F3: if θ and ϕ are mutually exclusive then $p(\theta \vee \phi) = p(\theta) + p(\phi)$.

Define *conditional (finitary) probability* by:

FC: if $p(\theta) > 0$ then $p(\phi|\theta) = p(\theta \wedge \phi)/p(\theta)$.

Then a finitary probability measure on $\mathcal{SL}$ determines a unique probability measure on the infinitary extension $\mathcal{S}^\infty\mathcal{L}$.

The reasoning behind this fact is as follows. $\mathcal{SL}/\equiv$ is isomorphic to the field of cylinders of truth functions considered above. A finitary probability measure on $\mathcal{SL}$ then induces a finitely-additive probability measure on the field of cylinders. But any such measure on the field of cylinders must be countably additive,[21] and can therefore be uniquely extended to a (countably additive) probability measure on the generated σ-field $\mathcal{F}$.[22] By the isomorphism between $\mathcal{F}$ and $\mathcal{S}^\infty\mathcal{L}/\equiv$ this corresponds to a probability measure in the infinitary system.

One could of course define probability over first-order predicate sentences. But if we define existential and universal quantification in terms of denumerable disjunction and conjunction respectively, and the first-order language has countably many atomic sentences, then for the purposes of defining probability the first-order language can be thought of as an infinitary propositional language over the atomic sentences. Thus in this situation the extra

[21] [Billingsley, 1979] theorem 2.3.

[22] [Billingsley, 1979] theorem 3.1.

expressive power of the predicate calculus is linguistic rather than probabilistic. Coupling this reduction with the reduction to a finitary propositional language, we can uniquely extend a finitary probability measure over the finitary propositional language to a probability measure over the first-order language.[23]

In sum, an infinitary propositional system is a useful formalism, not only because of the link with the mathematical definition of probability, but also because it acts as a good half-way house between more familiar finitary propositional systems and first-order predicate systems.

2.1 Partial entailment

Having seen how probability can be defined on sentences, we are now in a position to examine the concept of a probability logic which generalises or modifies classical logic. Such a system can be classified according to its interpretation of the logical implication operator $\models$. Perhaps the most obvious thing to try first is a generalisation of entailment $\models$ to partial entailment $\models_x$, where a set Θ of sentences partially entails sentence ϕ to degree x, $\Theta \models_x \phi$, if and only if $p(\phi | \bigwedge \Theta) = x$. Under such a view classical entailment is the case where $x = 1$. If Θ is empty we get a concept of degree of logical truth or degree of truth which corresponds to unconditional probability. There have been some well-known proponents of this kind of view, as we shall see now. In the following subsection we will examine the viability of the partial entailment approach.

One may be able to attribute the germs of such a logical approach to probability to some of the pioneers of probability — for example Boole permitted probabilities defined on propositions, although his logical foundations of probability really amounted to a mathematical calculus of probability derived logically, rather than a generalisation of logic.[24] However it wasn't until the developments and interest in formal logic at the end of the 19th and beginning of the 20th centuries that the view of probability as logic began in earnest.

Frege, Peano and Russell's contributions to formal logic motivated Łukasiewicz' innovative theory.[25] Instead of defining probability over sentences as we do above, Łukasiewicz defines probability over *indefinite propositions*, formulae which contain free variables. An indefinite proposition is true or false if true or false respectively for all substitutions. Further, '*By the truth value of an indefinite proposition I mean the ratio between the number of values of the variables for which the proposition yields true judgements*

[23] See [Gaifman, 1964], [Scott and Kraus, 1966], [Paris, 1994] chapter 11, or [Roeper and Leblanc, 1999] chapter 5 for further details concerning such reductions.

[24] See [Boole, 1854], [Boole, 1854b], [Boole, 1854c].

[25] [Łukasiewicz, 1913].

and the total number of values of the variables'.[26] Relative truth value is defined as conditional probability. Thus partial truth and partial entailment are given a specific interpretation.

Łukasiewicz rejects probability over events as being too restrictive in that it represents only the single-case. He distinguishes subjective and objective probability, but finds both interpretations unsatisfactory, subjective probability because it is too psychologistic, too subjective, and beliefs are unmeasurable (this was before the betting set-up had been introduced), and objective probability because determinism renders it redundant (this was before quantum mechanics) and because the principle of the excluded middle states that a proposition is objectively true or false at every time, precluding objective partial truth. Instead Łukasiewicz' interpretation is intended to be a coherent explication of Laplace's notoriously problematic *principle of indifference*, which defines a probability to be the ratio of the number of cases favourable to an event to the total number of possible cases, if we are indifferent as to which case will occur.[27] As he says,

> The interpretation of the essence of probability presented here might be called the *logical* theory of probability. According to this viewpoint, probability is only a property of propositions, i.e., of logical entities, and its explanation requires neither psychic processes nor the assumption of objective possibility. *Probability, as a purely logical concept, is a creative construction of the human mind, an instrument invented for the purpose of mastering those facts which cannot be interpreted by universally true judgements (laws of nature).*[28]

Keynes was another key player in the partial entailment tradition. He argued that probability generalises logic, measuring the degree to which an argument is conclusive. However he also allowed a subjective interpretation to his logical view of probability. In effect he proposed a probabilistic logic of rational belief.[29] Jeffreys too thought of probability as a generalisation of deductive logic, expressing support for an inference, given data.[30] In this respect his probability theory was a formalisation of inductive logic.[31]

Hempel closely studied this inductive relationship between evidence and hypothesis, deriving a qualitative logic of confirmation with a well-defined syntax and semantics: 'Confirmation as here conceived is a logical relationship between sentences, just as logical consequence is'.[32] Carnap rendered

[26] [Łukasiewicz, 1913] pg. 17.
[27] [Laplace, 1814].
[28] [Łukasiewicz, 1913] pg. 38.
[29] See [Keynes, 1921] section 1.1.
[30] [Jeffreys, 1931] section 2.0.
[31] [Jeffreys, 1939] section 1.2.
[32] [Hempel, 1945] pg. 24.

Hempel's theory quantitative by bringing probability into the logic. For Carnap probability was degree of confirmation. This was not cached out in terms of frequency (which he thought to be a valuable concept but quite different) or subjective degrees of belief (which he argued are too psychologistic), but given a distinct logical interpretation. The issue of confirmation is 'a logical question because, once a hypothesis is formulated by h and any possible evidence by e ..., the problem whether and how much h is confirmed by e is to be answered by a logical analysis of h and e and their relations. This question is not a question of facts in the sense that factual knowledge is required to find the answer'.[33]

Jaynes explicitly adopted a Keynesian position, arguing in favour of a logical partial entailment approach together with a subjective interpretation.[34] According to Jaynes' own theory, background knowledge imparts a fixed degree of plausibility (formally a conditional probability) to a proposition, and it is through principles like the principle of indifference, Laplace's rule of succession, the maximum entropy principle and symmetry constraints that we can identify the correct probability. The maximum entropy principle, for example, says that one should assign probabilities over the sentences of a finite language $\{c_1, \ldots, c_N\}$ in such a way that the entropy $(-\sum_\alpha p(\alpha)\log p(\alpha)$, where the α range over the atomic states $\pm c_1 \wedge \ldots \wedge \pm c_N$, where $+c_i$ is c_i and $-c_i$ is $\neg c_i)$ is maximised subject to any constraints imposed by background knowledge.[35]

2.2 Uniqueness

In this section we shall see that the partial entailment view of probability logic is not an easy one to maintain. In the next section we will discuss an alternative approach.

There are inconclusive objections specific to particular approaches. Language dependence is a trap for theories which rely on the principle of indifference, although its generalisation, the maximum entropy principle does not appear to be subject to these problems.[36] Theories that allow a subjective degree of rational belief interpretation to probability as defined over sentences suffer from the problem of logical omniscience. Here an agent is assumed to give the same degree of belief to logically equivalent sentences, even though the equivalence may not be known — the agent must somehow know all logical facts in order to be rational, which is rather a tall order, especially considering the fact that many difficult unsolved problems

[33] [Carnap, 1950] p. 20.

[34] [Jaynes, 1998] chapter 1.

[35] See [Paris, 1994] for justifications of the maximum entropy principle and an idea of the logical issues that surround the Keynes-Jaynes type of position.

[36] See [Paris, 1994], [Paris and Vencovská, 1997], [Paris, 1999].

in mathematics are questions of logical implication or equivalence.[37] On the other hand it is possible to represent uncertainty about logical implication even if we do adopt a subjective interpretation, as we shall see in the final section.

However, there is also an important general problem. The partial entailment approach requires that the degree x of confirmation that the set Θ of sentences gives to sentence ϕ, be a logical fact, dependent only on Θ and ϕ. Given the probabilistic interpretation of $\models_x$ as conditional probability, and letting Θ and ϕ vary, this means that there is some unique, distinguished probability function p^* which determines degree of confirmation, $\Theta \models_x \phi \Leftrightarrow p^*(\phi|\Theta) = x$. However, there are good reasons to doubt whether such a function p^* can be uniquely determined.

Łukasiewicz recognised the uniqueness requirement: 'Although probability does not exist objectively, the probability calculus is not a science of subjective processes and has a thoroughly objective nature. Hence the essence of probability must be sought not in a relationship between propositions and psychic states, but in a relationship between propositions and objective facts.'[38] Keynes considered the degree of belief interpretation unnecessary in as much as uniqueness leaves no room for subjectivity.[39] Jeffreys was of a similar opinion: the degree of belief interpretation is an optional extra. As he says, 'If we like there is no harm in saying that a probability expresses a degree of reasonable belief.'[40] Carnap also realised that if rational degree of belief is uniquely determined then the subjective element is gratuitous and can be omitted.[41] He puts it thus:

> The characterisation of logic in terms of correct or rational or justified belief is just as right but not more enlightening than to say mineralogy tells us how to think correctly about minerals. The reference to thinking may just as well be dropped in both cases. Then we say simply: mineralogy makes statements about minerals, and logic makes statements about logical relations. The activity in any field of knowledge involves, of course, thinking. But this does not mean that thinking belongs to the subject matter of all fields. It belongs to the subject matter of psychology but not to that of logic any more than to that of mineralogy.[42]

Jaynes on the other hand was a strict subjectivist and he thought that it is wrong to think of subjective entities as objective features of the physical

[37] See [Williamson, 1999c].

[38] [Lukasiewicz, 1913] p. 37.

[39] [Keynes, 1921] section 1.2.

[40] [Jeffreys, 1931] p. 22.

[41] [Carnap, 1950] section 2.11.

[42] [Carnap, 1950] pp. 41-42.

world, even if they do not vary from individual to individual.[43] However he did recognise that the uniqueness requirement precludes room for subjective disagreement (note that Jaynes' rational agent X is a robot):

> When we apply probability theory as the normative extension of logic, our concern is not with the personal probabilities that different people might happen to have; but with the probabilities that they "ought to" have, in view of their information ....
> In other words, at the beginning of a problem our concern is not with anybody's personal opinions, but with specifying the *prior information* on which our robot's opinions are to be based, in the context of the current problem. The principles for assigning prior probabilities consistently by logical analysis of that prior information are for us an essential part of probability theory.[44]

And,

> Surely the most elementary requirement of consistency demands that two persons with the same relevant prior information should assign the same prior probabilities.[45]

Thus uniqueness is recognised as a condition for the partial entailment position.[46] However there are two types of problem with uniqueness.

First, there seem general situations where two agents' beliefs may be rational yet differ. There are statements whose truth depends on the agent's perspective, such as "I am Xenelda". "Bob is tall" is a vague and to some extent subjective statement. Carnap might dismiss such statements as unscientific, and he defended his position against other examples.[47] More seriously, there are doubts as to how and when the constraining principles that Jaynes appeals to should be applied. For example, the principle of indifference may be applied in different ways depending on how the equipossible outcomes are delineated, and the maximum entropy principle can be interpreted as saying that one should not take risks, in as much as one should accept bets that minimise worst-case expected loss, which may not always be a good strategy.[48]

[43] This he called the *mind projection fallacy*: see [Jaynes, 1990] and [Jaynes, 1998] chapter 2 and pages 218, 1614.

[44] [Jaynes, 1998] Appendix A p. 5.

[45] [Jaynes, 1968] 228.

[46] Note that those who adopt the logical approach (with the uniqueness requirement) together with a subjective interpretation of probability (with Bayesian conditionalisation) occasionally call themselves 'objective Bayesians'. There is a possible source of confusion here: this position is objective in the sense that rational belief does not vary from individual to individual, *not* in the sense that I have been using, where an interpretation is objective if it is directly physical, not to do with subjects and their beliefs.

[47] [Carnap, 1950] sections 46-47.

[48] See [Grünwald, 2000].

Second, even if these problems are overcome, further problems arise on infinite domains. One can formulate constraining principles over languages involving a finite number of propositional variables, but it is generally not possible to extend them to a denumerable language, in particular a predicate calculus. Carnap himself searched for a unique probability function representing degree of confirmation but ended up with a continuum of probability functions. We saw in the last section that he viewed confirmation as a logical matter, not factual. However, if a function must be chosen from Carnap's continuum of inductive methods, this must be done either factually through calibration (for instance, seeing in practice which function agrees with frequency or leads to the best long-term betting strategy), or subjectively in an arbitrary fashion. Others have found that intuitively plausible rational constraints contradict each other on the infinite domain.[49]

One can see that there can be no unique most rational probability function on an infinite domain for the following reason.[50] Suppose we have a denumerable sequence $\theta_1, \theta_2, \ldots$ of mutually exclusive and exhaustive sentences. Agent X has no information about any of these sentences — how should she set her beliefs? We might want to generalise the principle of indifference to claim that each sentence should be given equal degree of belief. However, this degree of belief must be zero, since if $p(\theta_i) = \varepsilon > 0$ for all i then $\sum_{i=1}^{\infty} p(\theta_i)$ diverges, but by countable additivity $\sum_{i=1}^{\infty} p(\theta_i) = p(\bigvee_{i=1}^{\infty} \theta_i) = 1$, since the θ_i are exhaustive. Then if $\varepsilon = 0$, we get $\sum_{i=1}^{\infty} p(\theta_i) = 0$ which also contradicts countable additivity. Therefore the principle of indifference does not generalise, and some θ_i and θ_j must be awarded different probabilities. Since there is no information about either of these sentences, a belief function that swaps their probabilities is equally rational. There is no unique most rational probability function.

Thus the uniqueness requirement poses difficulties for the partial entailment approach. However, there are several possible responses. Firstly there are many ways in which the approach can be altered to abandon such a requirement. Second one can retain the uniqueness requirement if one drops the subjective or logical interpretation in favour of an objective interpretation of probability. We shall consider these options in order.

2.3 Generalised partial entailment

Dropping the uniqueness requirement yields a new notion of entailment $\models_A$, where $\Theta \models_A \phi$ iff $A = \{x : x = p(\phi | \bigwedge \Theta) \text{ for some } p\}$.[51] This concept is a great deal weaker, for if Θ and ϕ are logically unrelated then $A = [0, 1]$ and

[49] [Wilmers *et al.*, 1999].

[50] [Williamson, 1999].

[51] [de Finetti, 1970] section 3.10 shows that A will always be a singleton or a closed interval.

the entailment relation says nothing. However, one can bolster the concept by limiting the probability functions under consideration when constructing the set A. Consider for example a subjective interpretation. Here the probability functions are rational belief functions of agents with knowledge Θ. Many subjectivists would accept the weak generalisation of partial entailment, but a subjectivist like Jaynes would apply extra principles, like maximum entropy, to narrow down the range of belief functions considered rational given the background knowledge. We saw in the last section that this need not narrow down A to a unique value, but it may nevertheless considerably strengthen the partial entailment relation.

Howson is a subjectivist of the former ilk. Degrees of belief are restricted to be probabilistic, but no further constraints are imposed (apart from the qualified use of Bayesian conditionalisation). Howson develops a logical approach to probability, based on the notion of consistency, which is defined by a set of axioms concerning fair bets. These are used to show that an assignment of degrees of belief is consistent if and only if it is the restriction of some probability measure.[52] We generate an entailment operator from this notion of consistency by appealing to Howson's identification of models with probability measures: an assignment of degrees of belief is consistent if and only if it has (is the restriction of) a model. Then assignment q entails assignment r iff all probability measures satisfying q, satisfy r. That is, q entails r iff r is the restriction of any probability measure p that extends q. If assignment q is of the form $q(\theta_1) = y_1, q(\theta_2) = y_2, \ldots$ and r is $r(\phi) = x$, this becomes: q entails r iff for all probability measures p, if $p(\theta_1) = y_1, p(\theta_2) = y_2, \ldots$ then $p(\phi) = x$. I shall call this formulation *probabilistic entailment*, and write $\theta_1/y_1, \theta_2/y_2, \ldots \models_x \phi$. Probabilistic entailment is just logical entailment via the axioms of probability. We then get the following connection with the earlier version of partial entailment: $\theta \models_x \phi \Leftrightarrow \forall p, p(\phi|\theta) = x \Leftrightarrow \forall p, p(\theta \wedge \phi) = xp(\theta) \Leftrightarrow \theta/y \models_{xy} \theta \wedge \phi$. Of course depending on the θ_i and ϕ there will often be no unique x such that $\theta_1/y_1, \theta_2/y_2, \ldots \models_x \phi$, in which case we can only say that $\theta_1/y_1, \theta_2/y_2, \ldots \models_A \phi$ for some set $A \subseteq [0,1]$. Then $\theta \models_A \phi \Leftrightarrow \theta/y \models_{Ay} \theta \wedge \phi$, where $Ay = \{xy : x \in A\}$. Thus Howson's proposal induces probabilistic entailment which is closely related to weak partial entailment.

Adams interprets probabilities as degrees of belief, but relies also on frequencies, since he argues that degrees of belief must approximate frequencies in order to be useful for practical reasoning.[53] Adams extends propositional logic by adding a new non-truth-functional conditional connective $\Rightarrow$, which can only link formulae which have the usual connectives in them (so one can have $\theta \Rightarrow (\phi \vee \psi)$ but not $\theta \Rightarrow (\phi \Rightarrow \psi)$)

[52] See [Howson, 2000]. The consistency condition is a normative constraint just as coherence is under the standard Dutch book foundations of subjective probability, which Howson rejects.

[53] [Adams, 1998] chapter 9 and appendix 1.

and for which $p(\theta \Rightarrow \phi) = p(\phi|\theta)$. Adams provides a probabilistic semantics: an inference is valid iff the uncertainty of its conclusion cannot exceed the sum of the uncertainties of its premises, where Adams defines uncertainty as $u(\theta) = 1 - p(\theta)$. It turns out[54] that an argument is valid in Adams' sense if and only if the premises entail the conclusion thus: $\forall\delta \in (0,1), \exists\varepsilon \in (0,1), [p(\theta_1) \geq 1-\varepsilon, \ldots, p(\theta_n) \geq 1-\varepsilon \Rightarrow p(\phi) \geq 1-\delta]$, which I will write as $\theta_1, \ldots, \theta_n \models_* \phi$ and call $*$-entailment. Note that $*$-entailment coincides with classical entailment on arguments which do not involve the connective $\Rightarrow$.[55]

2.4 Objective interpretations

While Howson and Adams are subjectivists, objective interpretations are also represented in the probability logic literature. Objective chance interpretations often appeal to a possible-world semantics for entailment. Above I equated the probability of a sentence with the probability of the set of truth functions that satisfy the sentence. One can think of this as the probability that the truth function corresponding to the actual world is in this set of truth functions, as the probability that the actual world is in the set of worlds at which the sentence is true, or as the proportion of possible worlds taken up with worlds at which the sentence is true. Nilsson's probability logic is based on this thought and employs the probabilistic entailment relation considered above.[56] One can also think of partial entailment in terms of possible worlds if one thinks of $p(\phi|\theta)$ as the proportion of θ-worlds in which ϕ is also true.[57] In this way one can retain the original concept of partial entailment, $\Theta \models_x \phi$ iff $x = p^*(\phi|\bigwedge\Theta)$, together with its uniqueness requirement, by interpreting p^* objectively.

Łukasiewicz gave the probability of an open sentence a logical interpretation, but it can also be given a frequency interpretation, in which the probability of $P(x)$ is the frequency of P. Coupling this with probabilities for sentences and a possible-worlds semantics, one can claim that the probability of a formula is the average frequency, with the average taken over classes of possible worlds and weighted by the chance of our world being in such a class. This is essentially the interpretation suggested by Loś and refined by Fenstad.[58]

[54] [Adams, 1998] section 7.1.

[55] $*$-entailment can also be used as a semantics for non-monotonic logic, if the numbers δ and ε are taken to be infinitesimals. See [Pearl, 1988] in this respect. Note however that non-standard probability measures lead to a range of conceptual problems, including failure of the archimedean property and difficulties to do with interpreting and measuring infinitesimal probabilities.

[56] [Nilsson, 1986].

[57] See [Adams, 1998] chapter 8.

[58] [Loś, 1963], [Fenstad, 1967].

Halpern adopts a similar approach, but instead of mixing the frequency distribution at a world with the chance distribution of the worlds, he makes a sharp distinction between the interpretation of the probabilities of open and closed formulae. The probabilities of open formulae are given a frequency interpretation, while the probabilities of sentences are given a distinct possible-world semantics.[59] In Halpern's theory, the probabilities of sentences are also thought of as subjective probabilities, but such an interpretation does not fit well with the possible-world semantics. It is much more intuitive and straightforward to apportion degrees of beliefs directly to sentences than to assess how likely our world is to be within a class of other possible worlds and then translate that probability to one over a sentence. Reformulating a problem in terms of possible worlds usually offers little or no extra insight for the price of a complicated and counter-intuitive ontology — possible worlds may be a necessary evil for chance theorists, but in my view they are best avoided by subjectivists.

3 PROBABILISTIC INFERENCE

I hope to have given an indication of the range of interpretations available, both to probability itself and to entailment in a probabilistic logic. In this section I will give a brief flavour of the approaches to inference using probability. In the following sections I shall argue for a reinterpretation of one of these approaches, and outline the ramifications for practical reasoning.

Probabilistic reasoning poses serious practical challenges. Consider the task of defining probability over a propositional language based on just a finite set of propositional variables $\mathcal{L} = \{c_1, \ldots, c_N\}$. To specify a probability measure (or finitary probability measure) over the sentences of this language, one must specify at least $2^N - 1$ probability values. For example, one can determine the probability measure from the values given to the 2^N atomic states, $p(\pm c_1 \wedge \ldots \wedge \pm c_N)$, and one of these is redundant since it can be determined by additivity from the others. Thus the *space complexity*, the amount of space required to store a probability measure, is exponential in the number of nodes. One can calculate the probabilities of other finitary sentences from the specified probabilities, but an exponential number of additions may be required. Therefore the *time complexity*, the amount of time required to perform a probabilistic inference, may severely restrict applications to practical reasoning. Likewise it may be practically difficult to check an assertion of entailment and even to check whether an assignment of probabilities over a set of statements is consistent (this is the *consistency problem*).

I will consider two types of strategy for overcoming these practical problems. First we shall look at probability logics and the approaches to infer-

[59] [Halpern, 1990].

ence they yield. After this we shall look at a technique from artificial intelligence, namely Bayesian networks. Later I will argue that the Bayesian network approach can be integrated into a logic.

3.1 Probability logics

From a logical point of view, the first step towards efficient probabilistic inference is to provide a proof theory that is sound and complete with respect to the chosen entailment relation. Adams gives a sound and complete proof theory for his logic.[60] Likewise, probabilistic entailment can be given a sound and complete proof theory but only for the above finite language, not for example for a first order language involving an unbounded domain.[61] However, a proof theory is only a start. Finding a proof of a conclusion from its premises is often a prohibitively complex problem, even if a proof exists.

Nilsson gives a geometrical method for bounding the probabilities of sentences. Recall that for generalised probabilistic entailment, $\theta_1/y_1, \theta_2/y_2, \ldots \models_A \phi$ iff $\forall p$, if $p(\theta_1) = y_1, p(\theta_2) = y_2, \ldots$ then $p(\phi) \in A$. Nilsson gives a matrix technique for calculating the parameter A given a finite set of premises, their probabilities and the conclusion. Linear programming can be used to solve the problem, but as Nilsson acknowledges, computational complexity remains a serious difficulty.[62]

Logic programming aims to simplify the proof problem for classical logic by focussing on a more restricted language than the full predicate calculus, employing resolution as a single rule of inference, and treating the failure to find a proof of a ground atomic formula as a proof of its negation.[63] One hope is that by adding probabilities to logic programs and making use of the computational advantages offered by logic programming, one might perform probabilistic inference efficiently. Probabilistic Horn abduction, for instance, can be used to find the most likely hypothesis that explains a set of evidence.[64]

Another such system, stochastic logic programming, was devised to represent the bias of a machine learning program over the hypothesis and instance spaces,[65] but the formalism may also be applied to first-order probabilistic reasoning.[66] The basic idea here is that proofs of a goal (an implicitly existentially quantified open formula) are given a loglinear distribution parameterised by features of those proofs, and the probability of an instantiation of the goal is defined as the sum of the probabilities of the proofs that

[60] [Adams, 1998] chapter 7.

[61] [Halpern, 1990].

[62] See [Nilsson, 1986]. Nilsson also suggests a solution to the consistency problem.

[63] See [Nilsson and Maluszyński, 1990].

[64] [Poole, 1992].

[65] [Muggleton, 1995].

[66] [Cussens, 1999], [Cussens, 2000].

result in that instantiation. In a sense this progresses Łukasiewicz' original aim of defining probabilities on formulae based on their purely logical characteristics.

3.2 Bayesian networks

A Bayesian network consists of a directed acyclic graph, or *dag*, G over the propositional variables $c_1, \ldots, c_N$ together with a set of specifying probability values $S = \{p(c_i|d_i) : d_i$ is a state of the parents of c_i in G, $i = 1, \ldots, N\}$.[67] Now, under an independence assumption,[68] namely that given its parent states d_i, each node c_i in G is probabilistically independent of any state s of other nodes not containing the descendants of $c_i, p(c_i|d_i \wedge s) = p(c_i|d_i)$, a Bayesian network suffices to determine a probability measure p over the sentences $\mathcal{SL}$ of $\mathcal{L}$. This is determined from the probabilities of the atomic states, which are given by the formula $p(\pm c_1 \wedge \ldots \wedge \pm c_N) = \prod_{i=1}^{N} p(\pm c_i|d_i)$ where the d_i are the parent states consistent with the atomic state. Furthermore, any probability distribution on $\mathcal{SL}$ can be represented by some Bayesian network.[69]

In particular, if the c_i are *causal* variables, and the graph G represents the causal relations amongst them, with an arrow from c_i to c_j if c_i is a direct cause of c_j, then it is thought that G will automatically be acyclic, and the independence assumption will be a valid assumption.[70]

Bayesian networks offer the following advantages. First, if one deals just with Bayesian networks then there is no consistency problem, because any allocation of specifying values in $[0, 1]$ is consistent in that it is the restriction of some probability measure. Second, depending on the structure of the graph, the number of specifying probabilities may be relatively small. For example, if the number of parents of a node is bounded then the number of probabilities required to specify the measure is linear in N. Third, also depending on the structure of the graph, propagation techniques[71] can be employed which allow the quick calculation of conditional probabilities of the form $p(c_i|\alpha)$, where α is a state of other nodes. For example, if the graph is singly-connected (there is at most one path between two nodes) then propagation can be carried out in time linear in N. Thus while in the worst case (which occurs when the graph is *complete*, that is when there is an arrow between any two nodes) there is nothing to be gained by using a Bayesian network, if the graph is of a suitable structure then both the space

[67] If c_i has no parents, $p(c_i|d_i)$ is just $p(c_i)$.

[68] The Bayesian network independence assumption is often called the *Markov* or *Causal Markov* condition.

[69] See [Pearl, 1988] or [Neapolitan, 1990] for more on the formal properties of Bayesian networks.

[70] [Pearl, 1988], [Neapolitan, 1990].

[71] [Neapolitan, 1990].

and time complexity will be dramatically reduced. It is generally presumed that causal graphs are simple enough to offer satisfactory reductions in complexity.

4 BAYESIAN NETWORKS PROPERLY INTERPRETED

In this section and the next I would like to give some of my views as to the future directions of probability logic. I believe that Bayesian networks offer significant potential in this area, but that their use hinges on their interpretation. For a detailed account of the arguments in this section, see [Williamson, 2000].

The key issue is the interpretation of probability. Thus far in the Bayesian network literature the probability measure is either interpreted objectively, or subjectively but under the presumption that the subjective probabilities are estimates of objective probabilities. In my view, neither of these interpretations are viable.

The objective interpretation fails because the independence assumption need not hold, when assumed of causality with respect to objective probability. The independence assumption requires that any probabilistic dependency amongst the nodes in the network be accounted for by the causal relations within the network. This can be expressed more precisely by the *principle of the common cause*: if c_i and c_j are probabilistically independent and neither is a cause of the other, then they have one or more common causes and the dependency is *screened off* by the states d of the common causes, $p(c_i|d \wedge c_j) = p(c_i|d)$. However, the principle of the common cause can be shown to fail, for the simple reason that causal connection is not the only way that nodes can be rendered probabilistically dependent. They may be dependent because they have related meaning, or they may be logically or mathematically related, they may be related by non-causal physical laws, or by local or boundary conditions, or they may even be probabilistically dependent purely by accident. In any of these eventualities the Bayesian network independence assumption can fail.

On the other hand, if we view the Bayesian network as the knowledge of an agent X, consisting of her picture of causality and her degrees of belief, and those degrees of belief are assumed to be estimates of objective probabilities, then there is a further reason why the independence assumption might fail. X may simply not know about all the causal variables relevant to those in her language, or she may not know of all the causal relations linking the variables already in her language. One can show that if X's causal graph is incomplete in this sense then the independence assumption is very unlikely to hold.

If we opt for an unconstrained subjective interpretation, losing the link between subjective and objective probabilities, then there is no reason *at* *all* to suppose the independence assumption might hold. For X may have any belief function she wishes, and only a few of those (a set of measure zero, to be precise) will satisfy any non-trivial independence assumption.

The only alternative, I claim, is to adopt a constrained subjective interpretation, along the lines of Jaynes' interpretation, by employing the maximum entropy principle.[72] The idea is that the components of the Bayesian network — the causal graph G and specified probabilities S — represent agent X's background knowledge, and the maximum entropy principle is used to derive the full probability measure p. The specified probabilities offer an immediate constraint on this process: the derived measure p must extend these specifiers. There is no immediate constraint imposed by the causal graph, and I propose the principle of *causal irrelevance* be used to invoke a constraint. This principle says that if a probability measure p is derived over $\mathcal{SL}$ via maximum entropy, next a new propositional variable c_{N+1} which is not a cause of any of the variables in $\mathcal{L}$ is added to $\mathcal{L}$ to give $\mathcal{L}^+$, and finally a new probability measure q over $\mathcal{SL}^+$ is derived, then q extends p, that is, the restriction $q|_{\mathcal{SL}} = p$. Intuitively there is no information that c_{N+1} is relevant to the other variables, so it should be considered irrelevant. It can be shown[73] that

E: under these constraints, the probability measure p derived by maximum entropy is the same as that derived by a Bayesian network under the independence assumption.

Thus if probability measure p is given this type of subjective interpretation, and is constrained by G and S, then the independence assumption holds and p can be represented by the Bayesian network on G and S.

5 NEW DIRECTIONS

5.1 *A causal logic*

Given the above interpretation of the components of a Bayesian network as background knowledge, we can define a logic, as follows. As before, we consider a language based on a set of *causal* propositional variables, and for practical reasons this set is finite. The components (causal graph G and probability specification S) of a Bayesian network together with a set of sentences Θ *partially entail* sentence ϕ to degree x, $G, S, \Theta \models_x \phi$, iff, given the information represented in G and S, an agent ought to believe ϕ to degree x, $p(\phi | \bigwedge \Theta) = x$. The maximum entropy principle is used to select the function p (uniquely, because the language is finite, and so we consider

[72] There are a finite number of propositional variables here, so we don't get problems of non-uniqueness.

[73] [Williamson, 2000].

degree x rather than set A of degrees) and therefore the degree x of partial entailment. The value x can be determined from the Bayesian network formed on the components G and S. Such a calculation can be viewed as a *proof* of the entailment, and by equivalence [*E*], Bayesian network proofs are sound and complete with respect to this concept of partial entailment. The space and time complexity of a proof depend on the structure of the graph G, and the time complexity also depends on the form of Θ and ϕ.

Recall that propagation techniques can be used to quickly calculate $p(a_i|\alpha)$ where α is a state of some of the other variables. These techniques can also be used to calculate $p(\phi|\theta)$, as follows. First note that $p(\phi|\alpha)$ can be broken down into propagations. Write ϕ in disjunctive normal form as $\bigvee_{j=1}^{m} \alpha_j$ where each α_j is of the form $\bigwedge_{k=1}^{r} \pm c_{i_k}$ and the α_j are mutually exclusive. Now $p(\alpha_j|\alpha) = 0$ if α_j is inconsistent with α, otherwise $p(\alpha_j|\alpha) = p(\pm c_{i_1}|\pm c_{i_2} \wedge \ldots \wedge \pm c_{i_r} \wedge \alpha) p(\pm c_{i_2}|\pm c_{i_3} \wedge \ldots \wedge \pm c_{i_r} \wedge \alpha) \ldots p(\pm c_{i_r}|\alpha)$. Then $p(\phi|\alpha) = \sum_{j=1}^{m} p(\alpha_j|\alpha)$ and can thus be decomposed into propagations. Now $p(\phi|\theta)$ can also be decomposed into propagations, for if we write θ in disjunctive normal form as $\bigvee_{j=1}^{m} \alpha_j$ where each α_j is of the form $\bigwedge_{k=1}^{r} \pm c_{i_k}$ and the α_j are mutually exclusive, then

$$p(\phi|\theta) = \frac{p(\theta|\phi)p(\phi)}{p(\theta)} = \frac{\sum_{j=1}^{m} p(\alpha_j|\phi)p(\phi)}{\sum_{j=1}^{m} p(\alpha_j)} = \frac{\sum_{j=1}^{m} p(\phi|\alpha_j)p(\alpha_j)}{\sum_{j=1}^{m} p(\alpha_j)}.$$

Thus propagation techniques can be used to perform a proof, but the more logically complex the sentences involved in the partial entailment, the greater the time complexity of the proof.

Given such a causal logic, one can ask what should happen when an agent's background knowledge does not take the form of the components G and S of a Bayesian network. In particular, how should one derive a probability measure if not all of the required probability specifiers are available?

Garside [1996] and Rhodes [1999] have provided answers to this question for the special cases in which the causal graph has a tree or inverted tree structure. However, while they appeal to the maximum entropy principle they do not allow the causal knowledge to constrain the maximum entropy solution in any way. I advocate the principle of causal irrelevance as an extra constraint, and this gives a different solution to the problem. I shall indicate here how the combination of causal irrelevance and maximum entropy can be used to shed light on the issue of incomplete background knowledge.

Suppose X's background knowledge consists of a causal dag G with each arrow $c_i \rightsquigarrow c_j$ labelled by a weight $w \in [-1, 1]$, and each root node labelled by its probability. If the weight is positive, this signifies that $p(c_j|c_i \wedge d) \geq p(c_j|\neg c_i \wedge d) + w$ for each state d of the other parents of c_j. If it is negative then $p(c_j|c_i \wedge d) \leq p(c_j|\neg c_i \wedge d) + w$ for each such state d (intuitively in this case c_i prevents c_j). Using the techniques involved in the proof of the equivalence result [*E*], one can show that the probability measure determined from such a labelled graph by the principles of causal irrelevance and maximum entropy is the same measure as that determined by the Bayesian network involving specifiers of the form $p(c_j|c_i \wedge d) = \frac{1+w}{2}, p(c_j|\neg c_i \wedge d) = \frac{1-w}{2}$. One may specialise further by considering labels of the form $w \in \{+,-\}$, with some $\varepsilon > 0$ given such that $c_i \leadsto^+ c_j$ implies $p(c_j|c_i \wedge d) \geq p(c_j|\neg c_i \wedge d) + \varepsilon$ and $c_i \leadsto^- c_j$ implies $p(c_j|c_i \wedge d) \leq p(c_j|\neg c_i \wedge d) - \varepsilon$. We then get an equivalence between such a structure and a Bayesian network in which if $c_i \leadsto^+ c_j$ then $p(c_j|\pm c_i \wedge d) = \frac{1\pm\varepsilon}{2}$, and similarly for prevention (and if root probabilities are omitted in the labelled graph they are specified to be $\frac{1}{2}$ in the Bayesian network). Thus we can define partial entailment $G, \Theta \models_x \phi$ where G is a labelled graph of one of the above varieties. This move not only extends the causal logic to situations where background knowledge is more limited, but can also significantly reduce the space complexity of the corresponding Bayesian network. In general, if we replace the values in the specification S by bounds on those values, we determine the specification S, and thereby a Bayesian network, by selecting the values within those bounds which maximise entropy. The task of finding the probability measure over the whole domain that maximises entropy is broken down into smaller and easier tasks, namely those of maximising the entropy of the individual specifying probabilities.

5.2 A proof logic

I claimed earlier that causal connection is only one type of link between variables, and that other links, such as logical relations, may also induce probabilistic dependencies. This observation motivates the hope that the Bayesian network toolkit may be applicable to these other types of links. I shall consider one such application here.

A logical proof of a sentence takes the form of an ordered list. Consider a propositional language with sentences $s, t, u, \ldots$ and the following proof of $s \to t, t \to u \vdash s \to u$, using the axiom system of [Mendelson, 1964] section 1.4:

1. $t \to u$ [hypothesis]
2. $s \to t$ [hypothesis]
3. $(s \to (t \to u)) \to ((s \to t) \to (s \to u))$ [axiom]
4. $(t \to u) \to (s \to (t \to u))$ [axiom]
5. $s \to (t \to u)$ [by 1, 4]
6. $(s \to t) \to (s \to u)$ [3, 5]
7. $s \to u$ [2, 6]

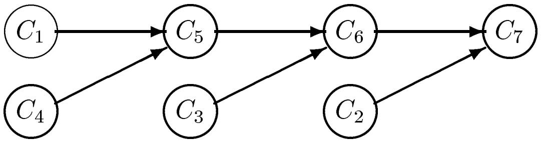

Figure 1. A proof dag.

The first important thing to note is that the ordering in a proof defines a directed acyclic graph. If we let c_i signify the sentence on line i, for $i = 1, \ldots, 7$, we get the dag in Figure 1.

By specifying degrees of belief in root nodes and conditional degrees of belief in other nodes given states of their parents, we can form a Bayesian network. These beliefs will depend on the meaning of the sentences, and in this example a specification might start like this: $S = \{p(c_1) = \frac{3}{4}, p(c_2) = \frac{1}{3}, p(c_3) = 1, p(c_4) = 1, p(c_5|c_1 \wedge c_4) = 1, p(c_5|c_1 \wedge \neg c_4) = \frac{1}{2}, \ldots\}$.

We can interpret the probability measure in various ways and if we assume that the probabilities are estimates of objective probabilities then, just as with causal Bayesian networks, we would not expect the independence assumption to hold unless (i) there are no dependencies due to non-logical relations amongst the variables and (ii) all the logical relations are included in the proof graph (but note that the graph will not necessarily remain acyclic if it includes all logical relations).

However the analogy with the causal case extends further, and if we adopt a subjective interpretation constrained by the maximum entropy principle, we would expect *proof irrelevance*, the analogue of the causal irrelevance condition, to hold. For if agent X learns of a new sentence c_8 that does not logically imply any of the others, her degrees of belief in the other sentences should intuitively not change.[74] Then the equivalence argument [E] can be used to justify equating X's belief function with the probability measure determined by the Bayesian network.

Finally, we can form a proof logic in the same way that we formed a causal logic above. While the parents of a node logically entail it, each node or set of nodes will also partially entail the others in the proof graph. The partial entailment relation can be proved, or the degree of partial entailment can be found, by using Bayesian network calculations. Thus two senses of proof are in play at once: a Bayesian network is used to prove a partial entailment, and the classical axiomatic method is used to prove the logical entailment on which the network is based.

[74] X learns of c_8 in the sense that she extends her language to include the new sentence, not in the sense of her learning the truth or falsity of c_8, which may well give her reason to change her other degrees of belief.

Why form beliefs about sentences in a proof? Because beliefs play a key role in proof planning. As Corfield demonstrates, subjective probability is important in mathematics: mathematicians regularly require degrees of belief in mathematical propositions, and these are apportioned according the mathematical and physical evidence available at the time.[75] The degrees of belief are instrumental in deciding whether to tackle a proof of a proposition (it may be worth tackling if the belief in the conclusion conditional on the premises is above a certain threshold) and in predicting from which area a possible proof will come (a conclusion may be more probable conditional on one set of premises or intermediary lemma than on another). To apply Bayesian proof networks to this type of problem we must recognise first that a parent of a node in the graph need not be one rule of inference away from the node. Just as a causal graph may represent causality on the macro-scale as well as the micro-scale, so too Bayesian proof networks may represent arguments involving large logical steps, as well as proofs like the one exemplified above. Second, just as in the causal case, an agent may not know of all the relevant logical factors or logical relations, and the proof network need not represent a complete proof — the proof graph will still be a dag. Thus the proof network includes the premises, conclusion and various other facts or conjectures which are considered relevant to the problem. The arrows in the network represent the flow of the proposed proof, and the probability specifiers represent the degrees of belief in each particular proposition, conditional on states of their parents. In this way a Bayesian proof network can be used to represent a belief function over a realistic mathematical problem, and to evaluate the conjectures and proposed proof paths.

Proof planning is not just important in mathematics. Whatever the domain and whatever the logic, if the proof theory relies on axiomatic deduction then finding a proof is likely to be a hard problem.[76] Automated theorem proving is now a large field of research, with applications ranging from software verification to robotics, but few systems tackle uncertainty in a fundamental way. Bayesian proof networks offer a practical opportunity for doing so. Moreover, proof planning is not the only application of such networks. Any domain of reasoning which requires the assessment of logically complex and connected variables may benefit from the approach, just as causal Bayesian networks are important for any kind of reasoning with causal variables, for example diagnosis, prediction and causal explanation.

[75] [Corfield, 2000].

[76] [Bundy, 1999], [Bundy, 2000].

6 SUMMARY

The choice of an appropriate interpretation of probability is of key importance, especially when it comes to combining probability with logic. Many logical approaches to probability ultimately fail from a philosophical point of view because of the uniqueness problem on infinite domains. However, probability logic can be sustained either by using a more general conception of partial entailment, by appealing to objective probability, or by sticking to finite domains. Bayesian networks take this latter approach and offer a way out of the practical problems that face a logic which incorporates probability, but again we must be careful about how we interpret probability. Probability for logic is best interpreted subjectively, using constraining principles like Jaynes' maximum entropy principle, if we wants to tap the power of Bayesian networks.

This power lies not only at the computational level, but also at the level of applications: Bayesian networks lead rather naturally to a causal logic and a proof logic. The causal logic may be applied to the problem posed at the beginning of this chapter by constructing a causal graph and probability specification (or bounds on these probability values), and using the associated Bayesian network to calculate the probability of X's house being burgled, given that her window is open and she hears an alarm. The proof logic may be applied to proof planning. Further extensions may be possible, and in the future we may expect some sort of integration between these extensions, either horizontally, where languages may involve mixtures of causal and logically complex variables for instance, or vertically where one logic is applied to a domain of causal variables and then a metalogic is applied to the logically complex sentences formed over that domain.

ACKNOWLEDGEMENTS

Thanks to David Corfield and Donald Gillies for useful comments, and to the Arts and Humanities Research Board for funding this research.

BIBLIOGRAPHY

[Adams, 1998] Ernest W. Adams. *A Primer of Probability Logic*, Stanford: CSLI Publications.

[Billingsley, 1979] Patrick Billingsley. *Probability and Measure*, John Wiley & Sons, third edition 1995.

[Boole, 1854] George Boole. *An Investigation of the Laws of Thought on which are Founded the Mathematical Theories of Logic and Probabilities*, London: Walton and Maberly.

[Boole, 1854b] George Boole. *Sketch of a Theory and Method of Probabilities Founded upon the Calculus of Logic*, in Studies in Logic and Probability, Watts & Co 1952, pages 141–166.

[Boole, 1854c] George Boole. *On a General Method in the Theory of Probabilities*, in Studies in Logic and Probability, Watts & Co 1952, pages 291–307.

[Bundy, 1999] Alan Bundy. A survey of automated deduction, Edinburgh Artificial Intelligence Research Paper 950.

[Bundy, 2000] Alan Bundy. A critique of proof planning, in [Kakas and Sadri, 2000].

[Carnap, 1950] Rudolf Carnap. *Logical Foundations of Probability*, Routledge and Kegan Paul Ltd.

[Corfield, 2000] David Corfield. Bayesianism in mathematics, in [Corfield and Williamson, 2000].

[Corfield and Williamson, 2000] David Corfield and Jon Williamson, eds. *Foundations of Bayesianism.* Kluwer Academic Publishers, 2001.

[Cussens, 1999] James Cussens. Loglinear models for first-order probabilistic reasoning. In *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, pages 126–133.

[Cussens, 2000] James Cussens. Stochastic logic programs: sampling, inference and applications. In *Proceedings of the 16th Annual Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, pages 115–122.

[Fenstad, 1967] J.E. Fenstad. Representations of probabilities defined on first order languages, in John N. Crossley, ed. *Sets, Models and Recursion Theory: Proceedings of the Summer School in Mathematical Logic and Tenth Logic Colloquium*, pages 156–172.

[Field, 1977] Hartry H. Field. Logic, meaning and conceptual role. *Journal of Philosophy*, **74**, 379–409, 1977.

[de Finetti, 1937] Bruno de Finetti. Foresight. Its logical laws, its subjective sources. In [Kyburg and Smokler, 1964], pages 93–158.

[de Finetti, 1970] Bruno de Finetti. 'Theory of probability', Wiley, 1974.

[Gaifman, 1964] H. Gaifman. Concerning measures in first order calculi. *Israel Journal of Mathematics* **2**, 1–18, 1964.

[Garside and Rhodes, 1996] Gerald R. Garside and Paul C. Rhodes. Computing marginal probabilities in causal multiway trees given incomplete information. *Knowledge-Based Systems*, **9**, 315–327, 1996.

[Garside *et al.*, 1999] G.R. Garside, P.C. Rhodes and D.E. Holmes. The efficient estimation of missing information in causal inverted multiway trees. *Knowledge-Based Systems*, **12**, 101–111, 1999.

[Grünwald, 2000] Peter Grünwald. Maximum entropy and the glasses you are looking through. In *Proceedings of the 16th conference of Uncertainty in Artificial Intelligence*, Stanford, 2000.

[Halpern, 1990] Joseph Y. Halpern. An analysis of first-order logics of probability. *Artificial Intelligence*, **46**, 311–350, 1990.

[Hempel, 1945] Carl G. Hempel. Studies in the logic of confirmation. In *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York: The Free Press 1965, 1970, pages 3–51.

[Howson, 1995] Colin Howson. Theories of probability. *British Journal for the Philosophy of Science*, **46**, 1–32, 1995.

[Howson, 2000] Colin Howson. The logical basis of uncertainty, in [Corfield and Williamson, 2000].

[Jaynes, 1968] E. Jaynes. Prior probabilities. In *IEEE Transactions Systems Science and Cybernetics*, *SSC-4*(3), 227, 1968.

[Jaynes, 1990] E.T. Jaynes. Probability theory as logic, in Paul F. Fougere, ed. *Maximum Entropy and Bayesian Methods*, Kluwer, 1990.

[Jaynes, 1998] E.T. Jaynes. Probability theory: the logic of science, http://bayes.wustl.edu/etj/prob.html.

[Jeffreys, 1931] Harold Jeffreys. *Scientific Inference*, Cambridge: Cambridge University Press, second edition 1957.

[Jeffreys, 1939] Harold Jeffreys. *Theory of Probability*, Oxford: Clarendon Press, third edition 1961.

[Kakas and Sadri, 2000] A. Kakas and F. Sadri. *Essays in Honour of Robert Kowalski*, to appear.

[Karp, 1964] C.R. Karp. *Languages with Expressions of Infinite Length*, North-Holland, 1964.

[Keynes, 1921] John Maynard Keynes. *A Treatise on Probability*, London: Macmillan, 1948.

[Kolmogorov, 1933] A.N. Kolmogorov. *The Foundations of the Theory of Probability*, New York: Chelsea Publishing Company, 1950.

[Kyburg and Smokler, 1964] H.E. Kyburg and H.E. Smokler, eds. *Studies in Subjective Probability*, New York: John Wiley, 1964.

[Laplace, 1814] Pierre Simon-Marquis de Laplace. *A Philosophical Essay on Probabilities*, New York: Dover, 1951.

[Lewis, 1980] David K. Lewis. A subjectivist's guide to objective chance, in [Lewis, 1986], pages 83–132, 1980.

[Lewis, 1986] David K. Lewis. *Philosophical Papers Volume II*, New York: Oxford University Press, 1986.

[Loś, 1963] J. Loś. Remarks on the foundations of probability. In *Proceedings of the 1962 International Congress of Mathematicians*, pages 225–229, 1963.

[Łukasiewicz, 1913] Jan Łukasiewicz. Logical foundations of probability theory. In L. Borkowski, ed. *Jan Łukasiewicz Selected Works*, Amsterdam: North-Holland 1970, pages 16–63.

[Mellor, 1971] David H. Mellor. *The Matter of Chance*, Cambridge: Cambridge University Press, 1971.

[Mendelson, 1964] Elliott Mendelson. *Introduction to Mathematical Logic*, Chapman and Hall, fourth edition 1997.

[von Mises, 1928] Richard von Mises. *Probability, Statistics and Truth*, Allen and Unwin, 2nd edition 1957.

[von Mises, 1964] Richard von Mises. *Mathematical Theory of Probability and Statistics*, Academic Press, 1964.

[Muggleton, 1995] Stephen Muggleton. Stochastic logic programs, in L. De Raedt, ed. *Advances in Inductive Logic Programming*, IOS Press, 1995.

[Neapolitan, 1990] Richard E. Neapolitan. *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*, New York: Wiley, 1990.

[Neapolitan, 1992] Richard E. Neapolitan. A limiting frequency approach to probability based on the weak law of large numbers. *Philosophy of Science*, **59**, 389–407, 1992.

[Nilsson, 1986] Nils J. Nilsson. Probabilistic logic, *Artificial Intelligence*, **28**, 71–87, 1986.

[Nilsson and Maluszyński, 1990] Ulf Nilsson and Jan Maluszyński. *Logic, Programming and Prolog*, Chichester: John Wiley andSons, second edition, 1995.

[Paris, 1994] Jeff Paris. *The Uncertain Reasoner's Companion*, Cambridge: Cambridge University Press, 1994.

[Paris, 1999] Jeff Paris. Common sense and maximum entropy, *Synthese*, **117**, 73–93, 1999.

[Paris and Vencovská, 1997] Jeff Paris and Alena Vencovská. In defense of the maximum entropy inference process, *International Journal of Automated Reasoning*, **17**, 77–103, 1997.

[Pearl, 1988] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Mateo, California: Morgan Kaufmann, 1988.

[Peirce, 1910] Charles Sanders Peirce. Notes on the doctrine of chances. In *Collected papers, Vol 2*, Harvard University Press 1932, page 404.

[von Plato, 1994] Jan von Plato. *Creating Modern Probability: Its Mathematics, Physics and Philosophy in Historical Perspective*, Cambridge: Cambridge University Press, 1994.

[Poole, 1992] D. Poole. Logic programming, abduction and probability. In *Proceedings, International Conference on Fifth Generation Computer Systems 1992*, vol 2, page 530, 1992.

[Popper, 1934] Karl R. Popper. *The Logic of Scientific Discovery*, London: Hutchinson, 1974.

[Popper, 1972] Karl R. Popper. *Objective Knowledge: An Evolutionary Approach*, Revised 1979, Oxford: Oxford University Press.

[Popper, 1983] Karl R. Popper. *Realism and the Aim of Science*, Hutchinson, 1983.

[Ramsey, 1926] Frank Plumpton Ramsey. Truth and probability. In [Kyburg and Smokler, 1964], pages 61–92, 1926.

[Roeper and Leblanc, 1999] P. Roeper and H. Leblanc. *Probability Theory and Probability Logic*, University of Toronto Press, 1999.

[Scott and Kraus, 1966] Dana Scott and Peter Kraus. Assigning probabilities to logical formulas. In Jaakko Hintikka and Patrick Suppes, eds. *Aspects of Inductive Logic*, Amsterdam: North-Holland, pages 219–264, 1966.

[Williamson, 1999] Jon Williamson. Countable additivity and subjective probability, *British Journal for the Philosophy of Science*, **50**(3), 401–416, 1999.

[Williamson, 1999b] Jon Williamson. The actual frequency interpretation of probability, philosophy.ai report pai_jw_99_b, http://www.kcl.ac.uk/philosophy.ai.

[Williamson, 1999c] Jon Williamson. Logical omniscience and rational belief, philosophy.ai report pai_jw_99_e, http://www.kcl.ac.uk/philosophy.ai.

[Williamson, 2000] Jon Williamson. Foundations for Bayesian networks. In [Corfield and Williamson, 2000], 2000.

[Wilmers *et al.*, 1999] G.M. Wilmers, M.J. Hill and J.B. Paris. Some observations on induction in predicate probabilistic reasoning, Department of Mathematics, Manchester University, 1999.