Journal of Logic, Language and Information **12:** 53–97, 2003.

Bayesianism and Language Change

JON WILLIAMSON
Department of Philosophy, King's College, London WC2R 2LS, U.K.
E-mail: jon.williamson@kcl.ac.uk

(Received 5 April 2002; in final form 25 July 2002)

Abstract. Bayesian probability is normally defined over a fixed language or event space. But in practice language is susceptible to change, and the question naturally arises as to how Bayesian degrees of belief should change as language changes. I argue here that this question poses a serious challenge to Bayesianism. The Bayesian may be able to meet this challenge however, and I outline a practical method for changing degrees of belief over changes in finite propositional languages.

Key words: Bayesian, Bayesian network, inductive logic, language change, maximum entropy

1. Introduction

The problem that I wish to address in this essay was identified by Imre Lakatos in his critical analysis of inductive logic:

> What is wrong with 'Bayesian conditionalisation'? Not only that it is '*atheoretical*' but that it is *acritical*. There is no way to discard the Initial Creative Act: the learning process is strictly confined to the initial prison of the language. Explanations that break languages and criticisms that break languages are impossible in this set-up (Lakatos, 1968: 347).

Colin Howson concurs:

> An objection, which in my opinion is a considerable one, to this procedure of representing his changes of belief is that it involves, as I remarked, the specification within a fixed language of his total possible future experience, and it commits him for all subsequent times to the way at some initial time he considered this range of possibilities as bearing on the set of events upon whose occurrence he will bet. This seems to me, as it has done to others, unrealistic (Howson, 1976: 296).

These passages relate to two quite distinct problems that beset Bayesianism. First, Bayesian conditionalisation requires that an agent always remain consistent with a prior probability distribution. Specifically, by Bayes' theorem the probability awarded to hypothesis h at time $t + 1$ is fixed by the evidence e arriving between times t and $t + 1$ and the prior probabilities of h and e and of e given h: $p_{t+1}(h) = p_t(h \mid e) = p_t(e \mid h)p_t(h)/p_t(e)$. However, the agent may decide that her prior p_t did not adequately assess the hypothesis or the evidence (perhaps

because she did not take sufficient notice of background knowledge)[⋆] or the relationship between the hypothesis and the evidence (perhaps she did not realise that the evidence follows logically from the hypothesis).[⋆⋆] Thus there may be good reasons to break out of the constraints imposed by a strict adherence to Bayesian conditionalisation.

The second problem is that Bayesian probability is normally defined on a fixed language or event space. Given this fixed framework, Bayesianism gives advice as to what degrees of belief to award sentences or events: having fixed a prior one should fix future degrees of belief by Bayesian conditionalisation. But in practice an agent's language often changes over time. There may be new sentences or events which were not even considered when formulating a prior, in which case Bayesian conditionalisation cannot be applied and Bayesianism fails to offer any guidance as to what degrees of belief to ascribe.[‡]

There are various possible solutions to the first problem. One strategy is to play down the role of Bayesian conditionalisation. One can accept that there are situations in which Bayesian conditionalisation is inappropriate, and allow other ways of updating beliefs.[‡‡] Another strategy is to play down the role of the prior. *Strict subjectivists* deny the coherence of physical interpretations of probability (frequency, propensity or chance interpretations), and consequently they reject any principle which asserts that physical probability should constrain an agent's degrees of belief. Such Bayesians often hold that prior beliefs are *washed out*, that is, as agents with different priors conditionalise on the same new evidence their belief functions converge, and consequently their priors have less of a bearing on their current beliefs.[¶] Hence for strict subjectivists the problem of having to remain consistent with a prior becomes less of an issue as time progresses.

The second problem has not been adequately addressed in the literature, as far as I know. It is this problem of language change that I shall consider here.

The problem of language change is particularly relevant today. This is because Bayesian theory is increasingly applied to artificial intelligence (AI) (see Pearl, 1988) and within AI the automated learning of new linguistic terms is an increas-

[⋆] Jaynes (1998) stresses the importance of ensuring that priors take background knowledge into account, and of correcting or reformulating a prior if it is realised that the prior does not adequately encode features of the background knowledge.

[⋆⋆] A Bayesian is usually presumed to be logically omniscient, and Earman (1992: 196) argues that belief changes that do not conform to Bayesian conditionalisation may be appropriate when this assumption fails.

[‡] I shall henceforth discuss only the case in which probability is defined over a language. An event-space framework can be treated analogously, as is shown in Williamson (2002a: §2).

[‡‡] As mentioned above, Jaynes and Earman follow this line by advocating a reassessment of priors. Howson claims that Bayesian conditionalisation should not be universally adopted because it can lead to inconsistencies (see Howson, 1997, 2001), and argues in Howson and Urbach (1989: §13.e) that beliefs may be updated by setting them to frequencies where they are known.

[¶] See de Finetti (1937) and Gaifman and Snir (1982) for convergence theorems, and Chapter 5 of Jaynes (1998) for a contrary view.

ingly important task.* The question now arises: how should the degrees of belief of an artificial agent change as its language changes?

Another key application of Bayesianism is within the philosophy of science, to confirmation theory (Howson and Urbach, 1989; Earman, 1992). In this context the problem of language change is crucial: competing scientific theories are often formulated in different scientific languages, and one must somehow bridge these languages in order to decide which theory is most confirmed by available evidence. Scientific theorising is often viewed as a special case of abduction, which may be thought of as the problem of formulating a plausible explanation of some given data (see Williamson, 2001a). Often one needs to change one's language in order to formulate a plausible explanation, either by adding new theoretical terms or by more radical reconceptualisations, yet one needs to evaluate the explanation in the light of the data which prompted it and any new data. Bayesianism is an important evaluatory framework – the most plausible hypothesis is usually considered to be that with maximum probability conditional on the data – hence Bayesianism must be extended to cope with changing language if it is to play a role in the abductive process, and scientific theorising in particular.**

These applications to AI and the philosophy of science pull Bayesianism in opposite directions. AI requires a formalism that is computationally practical, and this usually leads to a simple framework and strong assumptions – witness the theory of Bayesian networks applied to causal reasoning (Pearl, 1988; Neapolitan, 1990; Spirtes et al., 1993). But the philosophy of science often aims to be true to science as it is practised and this leads to an expressive linguistic formalism without restrictive assumptions: here probability is often used informally over natural language statements (Howson and Urbach, 1989; Earman, 1992) and may even be qualitative rather than quantitative (Polya, 1954: Chapter XV). But despite this methodological divergence the two disciplines are mutually supportive: the philosophy of science often motivates developments in AI and assesses AI assumptions, while AI systems can be used to empirically test philosophical accounts of scientific reasoning (see Thagard, 1988; Gillies, 1996; Williamson, 2001b). Consequently I shall pursue an integrated approach here. I shall first, in Part I, make some rather general comments on the problem of language change, arguing that an agent's choice of language expresses factual knowledge. This will motivate an

* There are various recent lines of development here. Concept learning is progressing at pace within statistical learning theory: See Vapnik (1995) and Cristianini and Shawe-Taylor (2000). New causes and effects are now automatically learned to improve the reliability of Bayesian networks: see Kwoh and Gillies (1996) and Binder et al. (1997). Multi-agent systems now evolve their own languages in order to communicate to solve problems (Jim and Giles, 2000). In the near future linguistic learning may also prove to be important in abductive logic programming (Kakas et al., 1998), inductive logic programming (Muggleton and de Raedt, 1994), and computational linguistics (Hausser, 1999), www.ling.ed.ac.uk/evolang2002/

** Note that in order to apply Bayesianism to science, one must also apply it to the mathematical theories on which the science depends (Corfield, 2001) and the comparison of mathematical theories in different mathematical languages is a significant problem in its own right (Kvasz, 2000).

AI-style solution to the problem in Part II, where I investigate the consequences of several assumptions within the restrictive linguistic framework of the propositional calculus.

Part I: THE GENERAL PROBLEM

2. Language Contains Implicit Knowledge

The problem of language change has rarely been discussed in the philosophy of science literature. But where it has been discussed, it has usually been in the context of an appeal to *language invariance*.[⋆] This is the claim that any assignment of prior probability should only depend on an agent's background knowledge, not on the underlying language. Or in the context of the current problem: an agent's probability function should not change when her language changes, unless she learns new facts at the same time.[⋆⋆] I shall argue, however, that *language contains implicit knowledge*. This creates a problem for the principle of language invariance, namely, language invariance is vacuous in the context of the language change problem. For whenever an agent's language changes she will simultaneously gain new knowledge, in which case language invariance offers no constraint on her new probability function. To get over this problem we will (in Section 7) replace the language invariance principle with a *conservativity* principle: when an agent's language changes her new degrees of belief should be as close as possible to her old degrees of belief, given her new knowledge. To apply this new principle we will need to choose an appropriate notion of closeness, make the implicit linguistic knowledge explicit, and specify how that knowledge constrains belief change. The formalities will be dealt with in Part II. For now I shall focus on the claim that language invariance cannot be applied naively.

There are two main ways that language represents knowledge. The choice of predicates in the language says something about those predicates themselves (Section 3) and about how the predicates relate to each other (Sections 4 and 5).

3. Goodman's New Problem of Induction

Nelson Goodman's new problem of induction shows us one way in which inductive inference is not language invariant. Goodman pointed out that some predicates (like

[⋆] See §G of the preface to the second edition of Carnap (1950), Carnap (1971: §§2.A.2-4 and 6.T6-1), Rosenkrantz (1977: §3.6), Forster (1995: §5), Halpern and Koller (1995), Jaynes (1998). Paris (1994, 1997) adopt a notion of language invariance that is weaker than that considered here; the 'representation independence' of Paris and Vencovská (1997) corresponds more closely to the concept of language invariance in this paper.

[⋆⋆] Strictly speaking, if the domain of the agent's probability function changes then her probability function changes. Thus a precise formulation of the language invariance principle must say something like: the probability of any sentence of the new language should be the same as the probability given to its translation into the old language, if such a translation exists, and if no new factual knowledge is gained in the transition between languages.

"green") are amenable to inductive generalisation while others (such as "grue:" green before time t and blue after t) are not (Goodman, 1954: §3.4). Predicates of the former variety are called *projectible* and often refer to what are called *natural kinds*. We tend to include projectible predicates in our natural and scientific languages in order to facilitate inductive reasoning. Hence a natural or scientific language implies certain facts about what the natural kinds are, and a change in language implies a corresponding change in background knowledge.

If languages get better at latching on to natural kinds as they evolve, then there is good reason to reject any straightforward application of the language invariance principle. Suppose for example that an agent's current language contains predicates "grue" and "bleen" rather than "green" and "blue." The agent believes "all emeralds are grue" to degree 0.99 (the changeover time t is some time in the future). But then her language changes, with "green" and "blue" replacing "grue" and "bleen." If, as I maintain, this change implies that the new predicates latch on to natural kinds better than the old predicates, then the change alone may warrant giving a lower value than 0.99 to "all emeralds are green before t and blue after t," the translation of the old sentence into the new language. On the other hand, the previous belief may be enough to warrant a value of 0.99 given to "all emeralds are green," even though the sentence in the old language "all emeralds are grue before t and bleen after t" may have had a much lower value.

The lesson to be learned here is that the principle of language invariance can only be applied if there is no change in knowledge as language changes. The principle should take *all* background knowledge into account, even implicit knowledge betokened by choice of language, such as that of natural kinds. This clearly limits the applicability of the principle.

I should mention that Howson and Urbach have cast Goodman's example in a different light (Howson and Urbach, 1989: §7.k). They argue that the new problem of induction is a case of underdetermination of theory by evidence, since for future t "all emeralds are green" and "all emeralds are grue" have the same empirical consequences up to the present. Howson and Urbach claim that it is the choice of prior that distinguishes the confirmation given to these two hypotheses: an agent may have given a higher prior probability to "all emeralds are green" in which case she will still believe that hypothesis to a greater degree after evidence is collected. None of this is incompatible with what I have said. However, there does appear to be a fact of the matter about which predicates are projectible – this is not just a subjective issue – and so an agent who has evidence that "green" is projectible while "grue" is not, will surely be *irrational* to give a higher prior probability to "all emeralds are grue." Bayesianism should reflect this: perhaps by invoking a constraint on priors to the effect that projectible concepts are awarded higher prior probability than non-projectible concepts. Now at first sight it appears that no agent can have evidence before t that "green" is projectible but "grue" is not. It seems at first sight that no constraint on priors which appeals to the syntax of expressions will be able to differentiate between "all emeralds are green" from "all emeralds

are grue." But, I claim, language evolves to latch on to projectible predicates, and so if one and not the other of the predicates "green" and "grue" occurs as a primitive predicate of the language, then that alone is evidence of its projectibility. I claim that there is a sortal division in the language between projectible and non-projectible concepts: predicates in the language are likely to be projectible, while ad hoc concepts like "green before t and blue after t" constructed from primitive linguistic predicates are unlikely to be projectible. This gives a syntactic basis on which a prior constraint could operate, and it is clear that language invariance is wildly inappropriate given such a prior constraint.*

4. The Principle of Indifference

Howson formulates a version of the principle of indifference whereby each model (up to isomorphism) of a formal language is given equal probability, and the probability of a sentence is the number of models satisfying that sentence multiplied by the probability of a model (Howson, 2001: 145). This formulation is not language invariant and Howson takes this fact as ground to reject the principle of indifference. But there is another way of looking at this. We can accept that choice of language conveys knowledge about which partition of models the principle of indifference should be applied to, in which case we should not expect applications of the principle of indifference to be language invariant. If we accept that reasoning by indifference is a mode of reasoning analogous to inductive generalisation then it will be the evolution of language, in the face of selective constraints generated by the quality of our decision-making, that decides the partition of indifference. In many cases where the principle of indifference can be applied in conflicting ways, there is one way which seems intuitively correct, or leads to better predictions (Jaynes, 1973). In such cases there is a fact of the matter as to which language leads to better inferences. In other cases different languages may lead to different belief assignments but the ensuing decisions may be of the same quality. In these cases it does not matter that the principle of indifference can be applied in different ways: agents with the same explicit background knowledge but different languages may adopt different belief functions yet remain equally rational.

One possible objection to this view is that *internal* application of the principle of indifference remains problematic. The problem is that within a language there may be two partitions of sentences over which we can apply the principle of indifference but which give conflicting conclusions. The answer, I think, is not to

* In other cases of underdetermination, simplicity is an issue. The problem is that given any hypothesis one can gerrymander a more complicated hypothesis with the same empirical consequences. Some Bayesians maintain that simpler hypotheses should be given higher priors or that they receive higher likelihoods (Rosenkrantz, 1977: Chapter 5; Howson and Urbach, 1989: §15.i.2). See also Sober (1975), Forster and Sober (1994), Forster (1995), and some notions of simplicity may be amenable to syntactic definition. But simplicity may itself be language-relative. Such constraints may also depend on the makeup of the agent under consideration: what is simple for a human agent is sometimes complicated for an artificial agent and vice versa.

apply the principle of indifference over partitions of sentences within a language, but to stick to *external* applications, exemplified by Howson's partition of models of the language. There is a grue-some analogy. Our language may have predicates "green" and "blue," but we may construct within our language the predicate "grue," by defining it in terms of green and blue. However, an application of inductive generalisation to both "grue" and "green" will give conflicting conclusions. If we accept that it is the language itself that contains the facts about projectibility then the solution is to avoid inductive generalisations on predicates constructed within the language.

5. Indirect Evidence

Choice of language can also imply the existence of relationships and connections amongst the referents of the linguistic terms. Lakatos argued that a language is a part of any scientific theory, since it implies connections:

> *The choice of a language for science implies a conjecture as to what is relevant for what, or what is connected, by natural necessity, with what.* For instance, in a language separating celestial from terrestrial phenomena, data about terrestrial projectiles may seem irrelevant to hypotheses about planetary motion. In the language of Newtonian dynamics they become relevant and change our betting quotients for planetary predictions (Lakatos, 1968: 362).

This is especially true of artificial languages, which are often constructed with a single application in mind. In an expert system for liver diagnosis, for example, most of the events or propositions referred to will be causally connected. This remains true even if the causal structure is uncertain or unknown: identifying a suitable set of variables that may be causally related is a crucial first step to identifying the causal connections that actually pertain. If new terms are added to the language of the expert system it is because they are causally related, or are likely to be causally related, to the terms already present. One should thus be cautious when applying any constraint on rational belief which renders variables probabilistically independent when no explicit connection has been asserted – the choice of language may implicitly connect the variables.

Lakatos also observed that introducing new terms into a language may change beliefs on the old terms:

> the problem of "indirect evidence" (I call "*indirect evidence relative to L in* L^*" an event which does not raise the probability of another event when both are described in L, but does so if they are expressed in a language L^*). In the examples given by Putnam and Nagel L was Carnap's "observational language" and L^* is the superseding theoretical language. But a situation of the same kind may occur whenever a theory is superseded by a new theory couched in a new language. Indirect evidence – a common phenomenon in the growth of knowledge – makes the degree of confirmation a function of L which, in turn changes as science progresses. Although growth of evidence *within* a fixed

> theoretical framework (the language L) leaves the chosen c-function unaltered, growth of the theoretical framework (introduction of a *new* language L^*) may change it radically.⋆

In general when language changes there is often implicit knowledge which both guides the ascription of degrees of belief over the new terms, and warrants a change in the beliefs over the old terms. We see this when we examine the ways in which language can change.

6. Types of Language Change

Perhaps the simplest form of language change occurs when the language expands to include new terms. There may be some feature of the world that one cannot describe in the current language, and so one needs to add a new propositional variable, constant, relation or function in order to do so. Typically the inadequacies of language are realised during abductive reasoning, that is, the search for an explanation or hypothesis. For instance, when Mendeleev developed the periodic classification of the elements, a theory was hypothesised which posited elements corresponding to each atomic weight – the referents of these new linguistic constants were only gradually discovered in the world. Similarly one may search for some causal explanation of a set of symptoms, find none in the current language, and so invent a syndrome which refers to the particular combination of symptoms, and invent a new causal term to signify whatever actually causes the syndrome. Further investigation then yields a clearer idea as to the properties of the new hypothesised cause. Note that new variables are often likely to be relevant to, and even indirect evidence for, old variables: on discovering a common cause of two symptoms, for example, one may judge the symptoms more dependent than previously thought.

Languages also contract. Non-referring or redundant terms are often eliminated: a new cause may be invoked to explain a syndrome, but then a cause in the old language may be found, leading to elimination of the new term. Alternatively a new cause may be found to refer, but to be irrelevant to the variables under consideration

⋆ Lakatos (1968: 363). Here Lakatos refers to Putnam's argument that Carnapian degree of confirmation cannot be defined on a language rich enough for science (Putnam, 1963), and Nagel's objection to Carnap (Nagel, 1963). Nagel observes that Carnap assumes a fixed language which is complete in the sense that it expresses all scientific terms, past, present and future (see Carnap, 1950: §18, 74–76). This is required because Carnap's degree of confirmation c^* depends on the number of terms in the language, and language variance is deemed counter-intuitive. Nagel argues that such an assumption is inappropriate, because it is doubtful as to whether we shall ever have a complete language for science, since theoretical (as opposed to observable) scientific terms undergo frequent changes. Nagel also argues that language variance is also a problem for Carnap in that a translation of one theory that involves polyadic primitives into an alternative language may change the number of primitives required, and therefore the confirmation function. (This is not so if the primitives are all monadic.) See Gillies (2001) for a further argument to the effect that Bayesianism requires fixity of the theoretical framework.

in the old language. Thus a variable may be eliminated if it is *not* indirect evidence. Similarly, if a relation is found always or never to obtain then it may be considered uninteresting and removed.

Of course language change can be more complicated. Languages may amalgamate, for instance. Alternatively there may be a non-trivial embedding of the old language into the new language. For example with the introduction of a distinction a propositional variable a may be replaced by b and c, in which case the transition from old to new language will be accompanied by the knowledge that $a \leftrightarrow b \vee c$. One interesting case is where the syntax of the language is the same, but the meaning of some of the terms changes. As Thomas Kuhn notes

> The need to change the meaning of established and familiar concepts is central to the revolutionary impact of Einstein's theory. Though subtler than the changes from geocentrism to heliocentrism, from phlogiston to oxygen, or from corpuscles to waves, the resulting conceptual transformation is no less decisively destructive of a previously established paradigm. We may even come to see it as a prototype for revolutionary reorientation in the sciences. Just because it did not involve the introduction of additional objects or concepts, the transition from Newtonian to Einsteinian mechanics illustrates with particular clarity the scientific revolution as a displacement of the conceptual network through which scientists view the world (Kuhn, 1962: 102).

Standard formulations of logic do not take into account the change in meaning of terms; thus a logical reconstruction of such cases may demand a change of syntax when meaning changes, so that, instead of a single mass term m being reinterpreted, Newtonian mass m^N is replaced by Einsteinian mass m^E.

According to Kuhn, two scientific theories may be incommensurable and it may be difficult to find grounds to prefer one over the other. Part of the problem is that it may be difficult for a proponent of one theory to translate the other theory into her own language.[⋆] This is a genuine problem for Bayesianism: how can an agent evaluate another theory if she cannot formulate that theory in her own language? Perhaps the only solution is to expand her language to formulate the new theory and update her beliefs on the basis of those links between the two languages of which she is aware. Thus if our agent has a belief function over the language of Newtonian mechanics and wants to evaluate special relativity, she could extend her language to include the language in which special relativity theory is formulated, and extend her belief function to this bridge language in the light of any constraints imposed by her knowledge of connections between the terms of the two languages.[⋆⋆]

⋆ Postscript to Kuhn (1962): pp. 202–204.

⋆⋆ One can interpret Kuhn's incommensurability thesis as the stronger claim that there is no common bridge language into which two theories can be translated. However, as Earman points out in Earman (1992: §8.2), there is little evidence for this thesis and in examples from the history of science it does always seem to be possible to contrive a (perhaps rather unnatural) overarching language.

7. Conservativity

The language invariance principle says that in the absence of any change in factual knowledge, an agent's belief function should not change as her language changes. I have argued that a change in language is accompanied by a corresponding implicit change in factual knowledge. This renders the language invariance principle inapplicable.

The *conservativity principle* is more practical. This says that when an agent's language changes her new degrees of belief should be as close as possible to her old degrees of belief, given her new knowledge. I will postpone a precise formulation of such a principle to Part II. In this section I will discuss the rationale behind conservativity from a general perspective.

Probability as degree of belief is usually justified by appealing to betting considerations. An agent's degree of belief in a sentence θ is interpreted as the *betting quotient* x she would give, were she to lose $(x - \tau(\theta))S$, where truth function $\tau = 1$ if θ is true and 0 if θ is false, and where S is an unknown stake which may be positive or negative. In order to avoid the possibility that stakes may be chosen that lead to loss whatever the true situation turns out to be, the agent's betting quotients must satisfy the axioms of probability (Ramsey, 1926; de Finetti, 1937; see also Williamson, 1999). Suppose the agent first adopts betting quotient x, and later changes her mind, adopting betting quotient y. Her loss function is then $(x - \tau(\theta))S_1 + (y - \tau(\theta))S_2$. Now it is possible to choose new stake S_2 so that the agent loses money whatever happens: if $S_2 > \max\{-x/yS_1, -(1-x)/(1-y)S_1\}$ then the loss will be positive, whatever the value of $\tau(\theta)$. This fact may be used to justify the claim that an agent should not change her degrees of belief unless she has good reason to. But suppose she does have good reason: she discovers that she will be irrational unless she chooses $y \in Y$, where Y is a closed subset of $[0, 1]$ such that $x \notin Y$. The agent's expected loss will be $y[(x-1)S_1 + (y-1)S_2] + (1-y)[xS_1 + yS_2] = (x-y)S_1$ which is clearly minimised if y is chosen to be the value in Y closest to x. Thus in order to minimise expected loss, the agent's new degree of belief must be as close as possible to her old degree of belief, subject to the constraints imposed by new knowledge. This gives a simple justification for conservativity.*

There is little doubt that humans are by nature conservative with respect to belief change.** As William James observes

> The individual has a stock of old opinions already, but he meets a new experience that puts them to a strain. Somebody contradicts them; or in a reflective

* Note that this justification assumes that minimisation of expected loss is an important goal – this may be disputed, especially considering the fact that expected loss is minimised just when expected gain (where gain is negative loss) is minimised. Note also that the situation becomes more complicated when we generalise from single degrees of belief to belief functions – see Section 10, where pointers to more comprehensive justifications are provided.

** In fact it appears we are often too conservative, holding on to beliefs even when we know them to be discredited – see Ross and Anderson (1982).

> moment he discovers that they contradict each other; or he hears of facts with which they are incompatible; or desires arise in him which they cease to satisfy. The result is an inward trouble to which his mind till then had been a stranger, and from which he seeks to escape by modifying his previous mass of opinions. He saves as much of it as he can, for in this matter of belief we are all extreme conservatives. ...
>
> New truth is always a go-between, a smoother over of transitions. It marries old opinion to new fact so as ever to show a minimum of jolt, a maximum of continuity (James, 1907: 148–149).

Conservativity has mainly been discussed in the context of propositional beliefs. However, much that has been said carries over to the Bayesian context of numerical degrees of belief, and it will be useful to examine the main positions.

There are a couple of blind alleys to be wary of. The first picks up on the fact that conservativity allows the possibility of two agents with the same evidence holding different beliefs but being equally rational. In the context of propositional beliefs this has been considered counter-intuitive.* But consider the same point in the context of numerical degrees of belief. Two agents start off with priors $p(\theta) = 1/4$ and $q(\theta) = 3/4$ respectively. They then both discover evidence that constrains rational degree of belief in θ to lie in $[1/3, 2/3]$. Changing their degrees of belief conservatively they arrive at the new values $p(\theta) = 1/3$ and $q(\theta) = 2/3$. These degrees are significantly different, yet based on the same evidence. However there should be nothing counter-intuitive here for a Bayesian. Bayesianism is built on the premise that different agents can hold different priors, and therefore different posteriors given the same evidence, yet both remain rational.

The second blind alley is the empirical justification of conservativity (Sklar, 1975: 387–388). Conservativity might be justified inductively if it could be shown that in the past minimal changes led more often to true theories than did extravagant changes of belief. This is a difficult line to take however, in view of the fact that we almost invariably change beliefs conservatively,** and, according to the pessimistic meta-induction (Laudan, 1981), our scientific theories are often proved wrong.

The more promising justifications of conservativity are typically pragmatic: it is a waste of time, energy and resources to continually change our beliefs for no reason, or to change them more than the minimum amount. William Lycan puts the point thus:

> Mother Nature would not want us to change our minds capriciously and for no reason. Any change of belief, like any change in social or political institution, exacts a price, by drawing on energy and resources. A habit of changing one's mind on a whim or otherwise gratuitously, like a habit of unrestrained social experimentation or a national disposition toward political coups or other sudden

* See Goldstick (1971). Sklar (1975) and Lycan (1988) hold a contrary view.

** Note that an agent does not always need to retain old beliefs in order to satisfy conservativity. Scientific revolutions may be considered to be instances of conservative belief change, where the minimal change in beliefs that is feasible in the light of new evidence is a revolutionary change.

> power and real estate grabs, would be inefficient and confusing; the instability it would create would be poorly suited to a creature whose need for cognitive organization in aid of sudden and streamlined action is great. (My wife points out that it does help, in the morning, not to have to reason your way to the bathroom.) (Lycan, 1988: 161).

Moreover, as Willard Van Orman Quine points out, we need to be conservative in order to explain new or unexpected phenomena within an existing framework:

> Familiarity of principle is what we are after when we contrive to "explain" new matters by old laws; e.g., when we devise a molecular hypothesis in order to bring the phenomena of heat, capillary attraction, and surface tension under the familiar old laws of mechanics. Familiarity of principle also figures when "unexpected observations" (i.e., ultimately, some undesirable conflict between sensory conditionings as mediated by the interanimation of sentences) prompt us to revise an old theory; the way in which familiarity of principle then figures is in favoring minimum revision.
>
> The helpfulness of familiarity of principle for the continuing activity of the creative imagination is a sort of paradox. Conservatism, a favoring of the inherited or invented conceptual scheme of one's own previous work, is at one the counsel of laziness and a strategy of discovery (Quine, 1960: 20).

Keith Lehrer takes the opposite view. He argues that conservativity inhibits discovery.

> The primary problem with this proposal is simply that it is a principle of epistemic conservatism, a precept to conserve accepted opinion. On some occasions, such a precept may provide good counsel, but often it will not. The overthrow of accepted opinion and the dictates of common sense are often essential to epistemic advance. Moreover, an epistemic adventurer may arrive at beliefs that are not only new and revelatory, but also better *justified* than those more comfortably held by others. The principle of the conservation of accepted opinion is a roadblock to inquiry, and, consequently, it must be removed (Lehrer, 1974: 184).

Of course, epistemic advances often require the overthrow of accepted opinion. But these advances occur because evidence in favour of new theories often renders old theories untenable for epistemic adventurers and conservatives alike. Lehrer's point misses the mark here for two reasons. The first is that he reads conservativity to entail that one should hold beliefs as close as possible to those of other people. This type of intersubjective agreement is only justifiable in special cases (Gillies, 1991). Indeed it seems quite plausible to hold that epistemic advances might be encouraged in the sciences if research councils fund individuals, each of whom are conservative with respect to their own beliefs, but who as a group hold a broad spectrum of incompatible beliefs. The second confusion in the above passage arises with the thought that the epistemic adventurer may be more justified than the conservative. No one argues that an agent should be conservative in the sense that she

ought to stick to her old beliefs in the face of evidence that justifies incompatible beliefs. The agent should change her beliefs to accommodate the new information, but change them only as much as is necessary. Thus Lehrer's arguments only succeed against notions of conservativity that few would be willing to uphold, and not the notion of conservativity that we are considering here.

It is wrong to think of conservativity in terms of justification. There is very little motivation for the assertion that a minimal change in beliefs is more justified than a large change in beliefs.* Justification has already done its work: given new knowledge certain belief states are justified; from those belief states (which are *all* justified) one ought to adopt the belief state which differs least from one's previous belief state. There is clearly no hope in claiming that justification determines that one should adopt that particular belief state. One should view the minimal change as most *rational* rather than most *justified*: it is for pragmatic reasons that we are and ought to be conservative.** Justification is to do with truth and evidential relations, whereas rationality must take both justification and pragmatics into account.

Gilbert Harman discusses conservativity from the point of view of belief revision. Harman distinguishes between *foundational* belief revision, where an agent keeps track of all the justifications of her beliefs and revises her beliefs according to this stock of knowledge, and *coherence* revision, where one forgets past justifications and assigns new beliefs on the basis of new information and the coherence of new beliefs with old beliefs (Harman, 1986: Chapter 4). Conservativity is then an important constraint for the coherence revision strategies: it allows one to choose a new belief state on the basis of the current state.

While Harman discusses belief revision in the context of propositional beliefs, the same distinction can be applied to numerical degrees of belief. Bayesian belief change is most naturally viewed as a coherence-based approach: Bayesian conditionalisation, for example, determines a new belief function from new evidence and the old function. Agents do not need to keep track of their justifications, and indeed it is of pragmatic advantage that they do not. A foundational approach to Bayesian belief change would require large amounts of space to store a database of all past evidence and justifications, and large amounts of time to maintain consistency of this database and to calculate a most rational belief function consistent

* As Lehrer points out elsewhere, "And the principle that, what is, is justified, is not a better principle of epistemology than of politics or morals" (Lehrer, 1978: 358). Christensen (2000) makes a similar point. Christensen puts forward the principle of *epistemic impartiality*, which says that an agent is not justified in adopting beliefs solely on the basis of their belonging to the agent's present belief state.

** It is for this reason that conservativity cannot help with the problem of underdetermination of theory by evidence. Sklar (1975: §3) argues that conservativity can be used to pick one among several equally justified hypotheses. But while conservativity can tell us what to do when we face underdetermination, the application of conservativity depends on there being underdetermination – if only one hypothesis is justified then we do not need conservativity to tell us what to do. Thus conservativity can in no way be thought of as a solution to the problem of underdetermination.

with the database. Thus coherence-based Bayesian updating offers what Harman calls "clutter avoidance:" the ability to avoid cluttering the mind with unimportant things (Harman, 1986: 41). It is no small matter to ensure that Bayesian degrees of belief can be stored efficiently[⋆] or that conservative Bayesian updating can be performed efficiently – these will be major concerns of Part II – but the potential is there with a coherence approach.[⋆⋆]

8. Prospects for a Solution

Lakatos again:

> Carnap tried his best to avoid any "language-dependence" of inductive logic. But he always assumed that the growth of science is in a sense cumulative: he held that one could stipulate that once the degree of confirmation of h given e has been established in a suitable "minimal language," no further argument can ever alter this value. But scientific change frequently implies change of language and change of language implies change in the corresponding c-values.
>
> This simple argument shows that Carnap's (implicit) "principle of minimal language" does not work. This principle of gradual construction of the c-function was meant to save the fascinating ideal of an eternal, absolutely valid, *a priori* inductive logic, the ideal of an inductive machine that, once programmed, may need an *extension of the original programming* but *no reprogramming*. Yet this ideal breaks down. The growth of science may destroy any particular confirmation theory: the inductive machine may have to be reprogrammed with each new major theoretical advance.
>
> Carnapians may retort that the revolutionary growth of science will produce a revolutionary growth of inductive logic. But how can inductive logic grow? How can we change our whole betting policy with respect to hypotheses expressed in a language L whenever a new theory couched in a new language L^* is proposed? (Lakatos, 1968: 363–364).

Most Bayesians – even those who accept objective constraints on priors – now reject Carnap's search for a unique, objective confirmation function, in favour of a subjective belief function relativised to an individual agent,[‡] yet Lakatos' questions at the end of this passage remain as important today as they were in 1968: we still do not know how degrees of belief should change as language changes.

[⋆] The choice of representation of probability function is crucial here – see Section 13.

[⋆⋆] Gärdenfors (1990: §3) picks up on the computational advantages that a coherence approach offers propositional belief revision. Indeed the AGM theory of belief revision that Gardenfors defends is a coherence theory. See also Rott (1999) on this point.

[‡] This is because there are clear cases in which there is more than one appropriate probability function – see Williamson (1999). Carnap himself at one time accepted that choice of belief function is to some extent subjective – see Hilpinen (1975: 337).

Earman maintains that there is no formal procedure for transforming the belief function in such circumstances:

> Indeed, the problem of the transition from Pr to Pr' can be thought of as no more and no less than the familiar Bayesian problem of assigning initial probabilities, only now with a new initial situation involving a new set of possibilities and a new information basis. But the problem we are now facing is quite unlike those allegedly solved by classical principles of indifference or modern variants thereof, such as E.T. Jaynes's maximum entropy principle, where it is assumed that we know nothing or very little about the possibilities in question. In typical cases the scientific community will possess a vast store of relevant experimental and theoretical information. Using that information to inform the redistribution of probabilities over the competing theories on the occasion of the introduction of the new theory or theories is a process that is, in the strict sense of the term, *a*rational: it cannot be accomplished by some neat formal rules or, to use Kuhn's term, by an algorithm. On the other hand, the process is far from being *ir*rational, since it is informed by reasons. But the reasons, as Kuhn has emphasized, come in the form of persuasions rather than proof. In Bayesian terms, the reasons are marshalled in the guise of plausibility arguments. The deployment of plausibility arguments is an art form for which there currently exists no taxonomy. And in view of the limitless variety of such arguments, it is unlikely that anything more than a superficial taxonomy can be developed (Earman, 1992: 197).

I am less sceptical. I think inroads can be made on the problem of language change, at least in a restrictive formal setting such as the propositional languages of Part II. Indeed unlike Earman I think that maximum entropy techniques can help us here. However there are cautionary lessons to be learned from the analysis of Part I. Language invariance will not help us because an agent's language contains implicit factual knowledge. This has two repercussions. Firstly if we are to save intuitions behind language invariance, we will have to generalise it to some form of conservativity principle. Secondly this transitional knowledge will have to be made explicit before the more general conservativity rule can be formally applied. Making the transitional knowledge explicit will in general be no mean feat – it is at this stage that insight and an awareness of subtleties of the particular domain come into play – but will clearly be a prerequisite of any formal analysis. Further, Kuhn's problem of incommensurability should lead us to look for a bridge language that encompasses the old and new languages. The relationships between the old and new terms in the bridge language may again be subtle and difficult to ascertain fully, but if knowledge of these relationships can be rendered explicit then the resulting formalisation will have normative value.

These then appear to be the key ingredients of a formal analysis: A conservativity principle generalising language invariance, transitional knowledge rendered explicit, and a bridge language involving both old and new terms.

Part II: PROPOSITIONAL LANGUAGES

9. Rational Assignments

We shall now look at the problem of language change from a more formal perspective. Our primary goal in Part II will be to produce a framework that can be implemented efficiently in an artificial agent.

Consider agent X whose rational belief function is p_0, a probability function on the sentences $\mathcal{SL}_0$ of propositional language $\mathcal{L}_0$.[⋆] Suppose X's language changes to $\mathcal{L}_1$ and K is an explicit formulation of *all* X's new knowledge gained in the transition from $\mathcal{L}_0$ to $\mathcal{L}_1$, including knowledge implied by choice of the new language. The key task is to define a new rational belief function p_1 over $\mathcal{SL}_1$.

We shall call K the *transitional knowledge*, and discuss its properties in more detail in Section 11. We shall call $\mathcal{L} = \mathcal{L}_0 \cup \mathcal{L}_1$ the *bridge language*, and $\mathcal{L}_+ = \mathcal{L} \backslash \mathcal{L}_0 = \mathcal{L}_1 \backslash \mathcal{L}_0$ the *additional language*. Note that if any of the variables in $\mathcal{L}_0$ change meaning in the transition to $\mathcal{L}_1$ (as in the case of the move from Newtonian to Einsteinian mass mentioned in Section 6) then the syntax should reflect this change by introducing new variables to correspond to the new meanings (thus the bridge language would contain a distinct variable for each type of mass). This framework allows us to achieve our key task by defining a rational belief function p on $\mathcal{L}$, given $\mathcal{L}_0$, p_0 and K, and then setting $p_1 = p|_{\mathcal{L}_1}$, the restriction of p to the sentences of $\mathcal{L}_1$.

Unfortunately, as pointed out in Section 1, Bayesian conditionalisation does not help us much here. The principle of Bayesian conditionalisation says that when X learns K she should set her new degrees of belief to her old degrees conditional on K, $p_1(\theta) = p_0(\theta \mid K)$, for each sentence θ of $\mathcal{L}_1$. This rule is fine when X's language does not change, but is only helpful in our context if $K \subseteq \mathcal{SL}_0$ and $\theta \in \mathcal{SL}_0$, since p_0 is only defined on the sentences of $\mathcal{L}_0$. Thus we require a more general way of specifying X's new belief function.

Let $\mathbb{P}_{\mathcal{L}}$ signify the set of probability functions defined on $\mathcal{SL}$. What we require is a way of transforming $p_0 \in \mathbb{P}_{\mathcal{L}_0}$ into $p \in \mathbb{P}_{\mathcal{L}}$ on the basis of transitional knowledge K. Define a *rational belief assignment* (or *assignment* for short) on $\mathcal{L}$ given p_0 and K, to be a function $\rho(\mathcal{L}, p_0, K)$ that selects a rational belief function $p \in \mathbb{P}_{\mathcal{L}}$, given rational $p_0 \in \mathbb{P}_{\mathcal{L}_0}$ and transitional knowledge K.

What form should ρ take? What makes a particular assignment $p \in \mathbb{P}_{\mathcal{L}}$ rational, given p_0 and K? This is the key issue we now face.

10. Conservative Assignments

In Sections 7 and 8 I mentioned that intuitions behind language invariance could be salvaged to some extent if we assume that rational belief should change as little as possible, as language and knowledge changes.

[⋆] See Paris (1994) for an introduction to probabilistic reasoning over propositional languages.

An assignment ρ is *conservative* if for each p_0 and K, $\rho(\mathcal{L}, p_0, K)$ is a function $p \in \mathbb{P}_{\mathcal{L}}$ satisfying K that is closest to p_0 according to some measure of distance between probability functions.

This yields a useful constraint on assignments once we specify a suitable distance function. Perhaps the most plausible measure of distance between probability functions on a finite language $\mathcal{L}_0$ is *cross entropy*:

$$d_{\mathcal{L}_0}(p, p_0) = \sum_{\alpha} p(\alpha) \log \frac{p(\alpha)}{p_0(\alpha)},$$

where the sum is over atomic states α of $\mathcal{L}_0$: if $\mathcal{L}_0 = \{c_1, \ldots, c_m\}$ then the *atomic states* $\mathcal{A}\mathcal{L}_0$ are the sentences of the form $\pm c_1 \wedge \ldots \wedge \pm c_m$. We shall use the standard conventions, justified by continuity arguments, that $0 \log 0/y = 0$ and $x \log x/0 = \infty$ for $x \neq 0$. Cross entropy is not a distance function in the usual mathematical sense, since it is not symmetric and does not satisfy the triangle inequality. However, we do have that $d_{\mathcal{L}_0}(p, p_0) \geq 0$ and $d_{\mathcal{L}_0}(p, p_0) = 0$ iff $p|_{\mathcal{L}_0} = p_0$,[⋆] which is enough for our purposes here.

There are several well-known arguments to the effect that a new belief function should minimise cross entropy relative to the old function, subject to constraints imposed by K.[⋆⋆] These arguments can be construed both as reason to employ conservative assignments in general and as reason to explicate the notion of conservativity via minimum cross entropy. We shall accept both these conclusions without further discussion, and we shall suppose for the rest of this paper that a conservative assignment minimises cross entropy between p and p_0 on $\mathcal{L}_0$. Since minimum cross entropy updating generalises Bayesian conditionalisation (Williams, 1980). the resulting conservative assignment will too.

Since we are dealing with finite domains here, we may plausibly require that *open-mindedness* be satisfied: $p(\theta) = 0$ iff θ is known (incontrovertibly) to be false.[‡] Open-mindedness is desirable for two technical reasons. First, since Bayesian conditionalisation can not update any zero degree of belief to a non-zero degree of belief, $p(\theta) = 0 \Rightarrow p(\theta \mid \phi) = 0$, it becomes necessary to save zero degrees of belief for incontrovertibly false sentences in order to apply Bayesian conditionalisation or its generalisations. Second, conditional probabilities are unconstrained if the probability of the condition is zero (conditional probabilities are subject to the constraint that $p(\theta \mid \phi)p(\phi) = p(\theta \wedge \phi)$, but this is vacuous when $p(\phi) = 0$). This may be appropriate when the condition ϕ is false, since according to the betting interpretation of rational belief conditional bets on false antecedents are called off, and so betting considerations do not constrain betting quotients in such cases.[‡‡] But if the condition is not known to be false, non-vacuous conditional probabilities may

⋆ See Paris (1994: proposition 8.5) for example.

⋆⋆ These are detailed in Paris (1994: 120–126).

‡ See Shimony (1955) and Chapter 7 of Paris (1994) for alternative formulations of this principle.

‡‡ Many doubt, moreover, the existence of objective truth conditions in counterfactual circumstances.

be useful. I may, for example, be inclined to give negligible degree of belief to the proposition that I will be run over when I next cross the road, yet I should not give it zero degree of belief because, given my background knowledge, I ought to give high degree of belief to an outcome of serious injury or worse, conditional on my being run over – that such conditional probabilities are non-vacuously constrained is vital for my decision making. For our purposes it is important to note that if open-mindedness initially holds, then any minimum cross entropy update will also satisfy open-mindedness, and the cross entropy distance between successive belief functions will never be infinite.

11. Compatible Transitional Knowledge

We need not assume here that K is set of sentences. Instead we shall view K more generally as a set of constraints on X's belief function p. (If $K \subseteq \mathcal{SL}$ then we can consider each sentence θ in K to be the constraint $p(\theta) = 1$.) Thus in effect we make two key assumptions about the transitional knowledge K: First that such knowledge can be made explicit at all, given that language change often involves implicit change in knowledge, and second that such knowledge can be made quantitative by articulating it as a set of well-defined probabilistic constraints. Most sciences require significant effort in identifying relevant constraints on a problem and then rendering these constraints quantitative, and it must be emphasized that such analysis is also required here.*

Let $\mathbb{K} \subseteq \mathbb{P}_{\mathcal{L}}$ be the set of probability functions on $\mathcal{SL}$ that satisfy K. We shall say K is *consistent* if there is some probability function that satisfies K, i.e., $\mathbb{K} \neq \emptyset$.

K is *compatible* with p_0 on $\mathcal{L}_0$ if there is a p on $\mathcal{L}$ satisfying K such that $p|_{\mathcal{L}_0} = p_0$. K is *compatible* on $\mathcal{L}_0$ if it is compatible with every p_0 on $\mathcal{L}_0$, i.e., if $\mathbb{K}|_{\mathcal{L}_0} = \mathbb{P}_{\mathcal{L}_0}$. Clearly K must be consistent to be compatible (with some p_0).

Here are some examples:

(i) $\mathcal{L}_0 = \{a, b\}$, $\mathcal{L} = \{a, b, c\}$, $K = \{a \leftrightarrow c, b \leftrightarrow c\}$. By the axioms of probability, if $p \in \mathbb{K}$ then $p(a) = p(c) = p(b)$ (Paris, 1994: proposition 2.1.c.) K is consistent since there exist probability functions that satisfy this requirement. But if $p_0(a) \neq p_0(b)$ then K is incompatible with p_0 on $\mathcal{L}_0$ – thus consistency does not imply compatibility. Intuitively c is indirect evidence for a and b, since the knowledge of c and its relationship with a and b makes a and b perfectly dependent.

(ii) $\mathcal{L}_0 = \{c_1, \ldots, c_m\}$, $\mathcal{L} = \{c_1, \ldots, c_m, c_{m+1}, \ldots, c_n\}$ and $K = \{p(\phi_1 \mid \theta_1) = u_1, \ldots, p(\phi_k \mid \theta_k) = u_k\}$ where the $u_i \in [0, 1]$, the $\theta_i \in \mathcal{SL}_0$ are disjoint (pairwise

* Frege made the point that the articulation and formalisation of a problem is often the most difficult task:

> I believe almost all errors made in inference to have their roots in the imperfection of concepts. Boole presupposes logically perfect concepts as ready to hand, and hence the most difficult part of the task as having been already discharged; he can then draw his inferences from the given assumptions by a mechanical process of computation (Frege, 1880: 34–35).

inconsistent) and the $\phi_i \in \mathcal{SL}_+$ (recall that $\mathcal{L}_+$ is the additional language $\mathcal{L}\backslash\mathcal{L}_0$). Then K is compatible on $\mathcal{L}_0$.

Proof. Given arbitrary p_0 on $\mathcal{L}_0$ we will define a probability function p that extends p_0 to $\mathcal{L}$ and satisfies K. Let the α run through the atomic states $\pm c_1 \wedge \ldots \wedge \pm c_n$ of $\mathcal{L}$. Let the β run through the atomic states $\pm c_1 \wedge \ldots \wedge \pm c_m$ of $\mathcal{L}_0$, and let β^α be $c_1^\alpha \wedge \ldots \wedge c_m^\alpha$, the atomic state β that is consistent with α. Similarly let the γ run through the atomic states of $\mathcal{L}_+ = \mathcal{L}\backslash\mathcal{L}_0$ and let γ^α be $c_{m+1}^\alpha \wedge \ldots \wedge c_n^\alpha$, the atomic state of $\mathcal{L}_+$ consistent with α. Let $|\theta_i| = |\{\beta : \beta \models \theta_i\}|$ and $|\phi_i| = |\{\gamma : \gamma \models \phi_i\}|$. Then define $p(\alpha) = q(\gamma^\alpha, \beta^\alpha) p_0(\beta^\alpha)$, where

$$q(\gamma, \beta) = \begin{cases} \frac{u_i}{|\phi_i|} : \ \beta \models \theta_i, \gamma \models \phi_i \\ \frac{1-u_i}{|\neg\phi_i|} : \ \beta \models \theta_i, \gamma \not\models \phi_i \\ \frac{1}{2^{n-m}} : \ \beta \not\models \bigvee_{i=1}^k \theta_i \end{cases}$$

(this is well defined by the disjointness of the θ_i). Now if $\beta \models \theta_i$ then

$$\begin{aligned} \sum_\gamma q(\gamma, \beta) &= \sum_{\gamma \models \phi_i} q(\gamma, \beta) + \sum_{\gamma \not\models \phi_i} q(\gamma, \beta) \\ &= u_i + (1 - u_i) \\ &= 1 \end{aligned}$$

and if $\beta \not\models \theta_i$ for any i then $\sum_\gamma q(\gamma, \beta) = \sum_\gamma 1/2^{n-m} = 1$. Hence $q(\gamma, \beta)$ can be interpreted as a conditional probability $p(\gamma \mid \beta)$ and p is a well-defined probability function extending p_0. Also,

$$\begin{aligned} p(\phi_i \mid \theta_i) &= \sum_{\beta \models \theta_i} q(\phi_i, \beta) p_0(\beta \mid \theta_i) \\ &= \sum_{\beta \models \theta_i} \left[\sum_{\gamma \models \phi_i} q(\gamma, \beta) \right] p_0(\beta \mid \theta_i) \\ &= \sum_{\beta \models \theta_i} [u_i] \, p_0(\beta \mid \theta_i) \\ &= u_i \sum_{\beta \models \theta_i} p_0(\beta \mid \theta_i) \\ &= u_i \, p_0(\theta_i \mid \theta_i) \\ &= u_i \end{aligned}$$

so p satisfies the constraints. □

This notion of compatibility provides us with two key properties. Clearly,

PROPOSITION 11.1 (Compatibility allows Extension). *If K is compatible with p_0 on $\mathcal{L}_0$ and ρ is a conservative assignment then $\rho(\mathcal{L}, p_0, K)$ extends p_0.*

We also have:

PROPOSITION 11.2 (Substitution of Equivalents).

- *Suppose $\theta \leftrightarrow \phi$ is in K, $\theta \in \mathcal{SL}_0, \phi \in \mathcal{SL}_1$.*
- *Let $\psi \in \mathcal{SL}$ and $\psi' \in \mathcal{SL}$ be $\psi[\phi/\theta]$, the result of substituting θ for ϕ in ψ.*
- ▸ *Then*

  1. *If $p \in \mathbb{K}$ then $p(\psi) = p(\psi')$.*
  2. *If $\psi \in \mathcal{SL}_1, \psi' \in \mathcal{SL}_0$, K is compatible with p_0 on $\mathcal{L}_0$, and p_1 is produced by a conservative assignment, then $p_1(\psi) = p_0(\psi')$.*

Proof. Let $\mathcal{L}' = \{c_1, \ldots, c_n\}$ be the smallest sublanguage of $\mathcal{L}$ containing all the above sentences (this is required in the case where $\mathcal{L}$ is infinite). Let $\alpha_1, \ldots, \alpha_{2^n}$ be the atomic states of $\mathcal{L}'$ (expressions of the form $\pm c_1 \wedge \ldots \wedge \pm c_n$.) Then any sentence $\sigma \in \mathcal{SL}'$ is logically equivalent to $\bigvee_{\alpha_i \models \sigma} \alpha_i$ and $p(\sigma) = \sum_{\alpha_i \models \sigma} p(\alpha_i)$. Let $\dot{\sigma} = \{\alpha_i : \alpha_i \models \sigma, p(\alpha_i) > 0\}$.

Part 1. Since p satisfies K, $p(\theta \leftrightarrow \phi) = 1$, an atom α_i only has positive probability if it satisfies $\theta \leftrightarrow \phi$. This implies that $\dot{\theta} = \dot{\phi}$, which in turn gives $\dot{\psi} = \dot{\psi}'$ as can be seen by induction on the complexity of ψ. Then $p(\psi) = \sum_{\alpha_i \in \dot{\psi}} p(\alpha_i) = \sum_{\alpha_i \in \dot{\psi}'} p(\alpha_i) = p(\psi')$, as required.

Part 2. p_1 is the restriction of p to $\mathcal{SL}_1$. Furthermore, by conservativity of the assignment and compatibility of K, p_0 is the restriction of p to $\mathcal{SL}_0$. Applying Part 1, $p_1(\psi) = p(\psi) = p(\psi') = p_0(\psi')$. □

12. Maximum Entropy

Minimising cross entropy will only constrain the new belief function p over $\mathcal{L}_0$. Suppose $\mathcal{L}_0 = \{c_1, \ldots, c_m\}$ and $\mathcal{L}_+ = \{c_{m+1}, \ldots, c_n\}$. Then ensuring that $d_{\mathcal{L}_0}(p, p_0)$ is minimised may fix the restriction $p|_{\mathcal{L}_0}$ of p to $\mathcal{L}_0$, but it will tell us nothing about p on $\mathcal{L}_+$. Thus we must look for a further constraint to choose an appropriate function p from all those functions equally close to p_0 on $\mathcal{L}_0$.

According to the *maximum entropy principle*, agent X should choose a function in $\mathbb{K}$ that maximises the entropy

$$H_{\mathcal{L}}(p) = -\sum_{\alpha \in \mathcal{A}\mathcal{L}} p(\alpha) \log p(\alpha),$$

where the sum is over atomic states α of $\mathcal{L}$. As with minimising cross-entropy, many of the justifications of the maximum entropy principle are well known,⋆ and I shall accept the principle without any further defence here.

⋆ See Jaynes (1998), Paris (1994) and Paris and Vencovská (2001) for justifications of the maximum entropy principle.

We thus have the following recipe for assigning p on $\mathcal{L}$ given p_0 and K: choose a function from $\{p \in \mathbb{K} : p \text{ minimises } d_{\mathcal{L}_0}(p, p_0)\}$ that maximises entropy $H_{\mathcal{L}}(p)$. We shall call this the *entropic* assignment, ρ_e, and we will focus on this assignment for the remainder of the paper.

In Section 7 I discussed the distinction between coherence and foundational approaches to belief change. The entropic assignment is conservative and thus coherence-based. But foundational assignments are also possible. For example, one could store all past background knowledge, evidence, and transitional knowledge that the agent has ever been exposed to, and whenever new knowledge enters this database – such as during language change – one could choose the agent's new belief function by maximising entropy subject to the constraints imposed by the database. As mentioned in Section 7, such a foundational approach is likely to incur enormous computational costs, and the reason for focusing on conservative assignments here is largely pragmatic.

An important special case occurs when $\mathbb{K}$ is a convex and compact set, for then minimising cross entropy fixes p uniquely over $\mathcal{L}_0$, and maximising entropy fixes p uniquely over $\mathcal{L}$. Thus the entropic assignment fixes a unique rational belief function in this case. This occurs quite often in practice, for example when the constraints in K are linear, i.e., of the form $\sum_{i=1}^{r} a_i\, p(\theta_i) = b$, for $\theta_i \in \mathcal{SL}$, $i = 1, \ldots r$ (Paris, 1994: proposition 6.1).

For the remainder of the paper we shall take $\mathcal{L}_0 = \{c_1, \ldots, c_m\}$, $\mathcal{L}_+ = \{c_{m+1}, \ldots, c_n\}$ and $\mathcal{L} = \{c_1, \ldots, c_n\} = \mathcal{L}_0 \cup \mathcal{L}_+$. Minimising cross entropy is maximising

$$\begin{aligned} -d_{\mathcal{L}_0}(p, p_0) &= -\sum_{\beta \in \mathcal{A}\mathcal{L}_0} p(\beta) \log \frac{p(\beta)}{p_0(\beta)} \\ &= -\sum_{\beta \in \mathcal{A}\mathcal{L}_0} p(\beta) \log p(\beta) + \sum_{\beta \in \mathcal{A}\mathcal{L}_0} p(\beta) \log p_0(\beta) \\ &= H_{\mathcal{L}_0}(p) + E \log p_0 \end{aligned}$$

over probability functions in $\mathbb{K}$, where E is the expectation with respect to p. Since $\log x$ is strictly increasing in x for $0 < x \leq 1$, the expectation of $\log x$ is maximised just when the expectation of x, judged according to future belief function p, is maximised. Thus minimising cross entropy can be thought of as a balance between maximising the entropy over $\mathcal{L}_0$ and maximising the expectation of current beliefs.

The next step, maximising entropy, requires maximising

$$\begin{aligned} H_{\mathcal{L}}(p) &= -\sum_{\beta \in \mathcal{A}\mathcal{L}_0, \gamma \in \mathcal{A}\mathcal{L}_+} p(\gamma \mid \beta) p(\beta) \log p(\gamma \mid \beta) p(\beta) \\ &= -\sum_{\beta \in \mathcal{A}\mathcal{L}_0, \gamma \in \mathcal{A}\mathcal{L}_+} p(\gamma \mid \beta) p(\beta) \log p(\gamma \mid \beta) - \sum_{\beta} p(\beta) \log p(\beta) \\ &= H_{\mathcal{L}_+ \mid \mathcal{L}_0}(p) + H_{\mathcal{L}_0}(p) \end{aligned}$$

over probability functions in $\mathbb{K}$. Now if $\mathbb{K}$ is convex and compact then the terms $p(\beta)$ are fixed by the cross entropy minimisation, and so entropy is maximised by maximising $-\sum_{\beta,\gamma} p(\gamma \mid \beta)p(\beta)\log p(\gamma \mid \beta)$ with respect to the parameters $p(\gamma \mid \beta)$.

13. Representing Probability Functions

In certain circumstances the entropic assignment may be implemented very efficiently. We shall explore these circumstances in the rest of the paper. The efficiency of the implementation will hinge on the way we choose to represent our probability functions. This section contains a brief overview of typical representations – those familiar with this material may skip to the end of the section.

A *representation* of a probability function p on $\mathcal{L}$ is a structure $\mathfrak{R} = (\mathcal{L}, G, S, A)$, where $\mathcal{L}$ is a language; G is some data structure; S is a set of probability specifiers – i.e., a set of statements of the form $p(\theta) = u$ where $\theta \in \mathcal{SL}$; and A is a set of assumptions linking G to the probability function p, such that $p : \mathcal{SL} \longrightarrow [0, 1]$ can be fully determined from $\mathfrak{R}$. Here are some examples, where $\mathcal{L}$ is in each case the finite propositional language $\{c_1, \ldots, c_n\}$:

(i) An *atomic representation* is of the form $\mathfrak{A} = (\mathcal{L}, \emptyset, S, \emptyset)$ where all probabilities of atomic sentences are specified, $S = \{p(\alpha) = u_\alpha : \alpha \in \mathcal{AL}\}$ such that $u_\alpha \in [0, 1]$ for each α and $\sum_\alpha u_\alpha = 1$. This is a representation of p because it determines p fully: for any sentence $\theta \in \mathcal{SL}$, $p(\theta) = \sum_{\alpha \in \mathcal{AL}, \alpha \models \theta} u_\alpha$. However, it is not a very compact representation of p, since S contains 2^n statements. One of these statements is redundant by the additivity of p, but without substantive assumptions (here $A = \emptyset$) representing p requires a set S of specifiers whose size is exponential in n. Other types of representation trade off size for assumptions in A.

(ii) A *Prospector representation* $\mathfrak{P} = (\mathcal{L}, G, S, A)$ is defined as follows. The variables in $G \subseteq \mathcal{L}$ are called *hypotheses*, while all other variables in $\mathcal{L}$ are *evidence* variables. The probability of each variable is specified together with the probability of each hypothesis given each evidence variable, $S = \{p(e_i) = u_i, p(h_j) = v_j, p(h_j \mid e_i) = w_{ij} : h_j \in G, e_i \in \mathcal{L}\backslash G\}$ where $u_i, v_j, w_{ij} \in [0, 1]$ and $\sum_j v_j = 1$. Finally in A it is assumed that the hypotheses in G are mutually exclusive and exhaustive and that the evidence variables are probabilistically independent conditional on the truth or falsity of a hypothesis, $p(e_1^\alpha, \ldots, e_k^\alpha \mid h^\alpha) = \prod_{i=1}^k p(e_i^\alpha \mid h^\alpha)$ for each state α of $\mathcal{L}$, and each hypothesis $h \in G$, where $e_1, \ldots, e_k$ are all the evidence variables. This is a representation of probability function p defined by $p(\alpha) = 0$ if zero, two or more hypotheses are true in α, and $p(\alpha) = v_j^{1-k} \prod_{\alpha \models e_i} u_i w_{ij} \prod_{\alpha \models \neg e_i} (v_j - u_i w_{ij})$ otherwise, where h_j is the hypothesis true in α, and then for $\theta \in \mathcal{SL}$, $p(\theta)$ can be defined as in (i). A prospector representation offers a very compact representation of p, since S contains at most $n^2/4 + n$ specifiers, and it can be used to determine $p(h_j \mid s)/p(\neg h_j \mid s)$ very efficiently, where s is some state of evidence variables. In fact such a representation

was used for probabilistic reasoning in Prospector, an expert system for geological prospecting in hard-rock mineral exploration. However, the assumptions made by a Prospector network are extremely strong: it can be shown that the assumptions imply that the evidence variables must all be probabilistically independent, and that each hypothesis is independent of all but one, or all, of the evidence variables.* Hence a more practical network might achieve a better balance between size and strength of assumptions than the extremes of either an atomic or a Prospector network.

(iii) A *Markov network* $\mathfrak{M} = (\mathcal{L}, G, S, A)$ may be defined thus. G is an undirected graph, whose nodes are the propositional variables in $\mathcal{L}$. The specification S is determined according to the following recipe. First triangulate G (in a *triangulated graph* every cycle of length four or more has a chord). Let $C_1, \ldots, C_k$ be the cliques of the triangulated graph (a *clique* is a maximal complete subgraph), ordered according to increasing values of $\max_{c_j \in C_i} j$. Form a *join tree* by connecting each clique C_i to a predecessor sharing the highest number of vertices with C_i. Let E_i be the sets of these vertices shared between C_i and its predecessor in the join tree, for $i = 1, \ldots, k$. Then specify in S the probabilities of all states of C_i conditional on all states of E_i, $S = \{p(C_i^\beta \mid E_i^\beta) = u_{i\beta} : \beta \in \mathcal{A}C_i\}$ where each $u_{i\beta} \in (0, 1)$ and for each state γ of E_i, $\sum_{\beta \sim \gamma} u_{i\beta} = 1$, the notation $\beta \sim \gamma$ signifying that β is consistent with γ. The assumptions A are that p is strictly positive, and that conditional on its boundary in G, each node is probabilistically independent of any set of other nodes (the *boundary* of a node in an undirected graph is the set of nodes adjacent to it). A set of variables that renders a variable independent of other variables is known as its *Markov blanket*. Under these assumptions the Markov network can be used to generate p since for each $\alpha \in \mathcal{A}\mathcal{L}$, $p(\alpha) = \prod_{i=1}^{k} u_{i\beta}$ where each β is consistent with α. *Propagation techniques* have been developed for calculating certain marginal probabilities from Markov networks.** The size of a Markov network and the speed of propagation calculations depend largely on the sizes of the cliques in the join tree: in the worst case both are exponential in n, but if the graph is sparse, space and time complexities can be dramatically reduced.‡

(iv) A *Bayesian network* $\mathfrak{B} = (\mathcal{L}, G, S, A)$ contains a directed acyclic graph G involving the propositional variables of $\mathcal{L}$ as nodes, and a probability specification S in which the probability of each node given each state of its parents in G is specified. A is an independence assumption, often called the *Markov condition*, which stipulates that the parents of each node in G render the node independent from any set of its non-descendants. This independence assumption ensures that a Bayesian network determines a full probability function, $p(\alpha) = \prod_{i=1}^{n} p(c_i^\alpha \mid d_i^\alpha)$ where d_i^α is the state of the parents of c_i consistent with α.‡‡ Thus one can use a

* See Paris (1994: Chapter 9) for details concerning the facts in this example.

** For more on Markov networks, see Pearl (1988: §3.2), Cowell et al. (1999: Chapter 5), and Lauritzen and Spiegelhalter (1988).

‡ See Neapolitan (1990: §7.6) on this point.

‡‡ See Pearl (1988: §3.3), and Neapolitan (1990: §5.3).

Bayesian network to represent a given probability function p on $\mathcal{L}$ by choosing a graph G with respect to which p satisfies the independence assumption and specifying $p(c_i \mid d_i^\alpha)$ in S for each node c_i and state d_i^α of its parents. The advantage of such a representation is computational: roughly speaking the sparser the graph, the smaller the amount of storage space the network takes up, and the quicker it is to calculate required probabilities from the Bayesian network. For example, if the maximum number of parents of nodes in graphs is bounded and graphs are singly-connected then the space complexity of Bayesian networks, and the time complexity for calculating $p(c_i \mid s)$ where s is a state of nodes, are both linear in n, whereas specifying each $p(\alpha)$ directly would lead to 2^n specifiers.⋆ Note that the graph G of a Bayesian network can be transformed into the graph of a Markov network by forming the *moral graph* of G by *marrying* the parents of each node (i.e., linking each pair of parents with an edge if they are unlinked) and replacing all arrows with undirected edges. By triangulating this graph we can achieve a Markov network representation of a strictly positive probability function p.

The strategy of the remainder of the paper will be to use a Bayesian network representation to implement the entropic assignment efficiently. We shall see in the next section that in certain circumstances a probability function p determined by the entropic assignment satisfies a number of conditional independence relationships that can be determined in advance. This allows us to construct the graph in a Bayesian network representation of p.⋆⋆ Moreover, using the entropic assignment to work out the corresponding specifiers in the Bayesian network representation is considerably easier than determining a direct representation of p. An algorithm for constructing a Bayesian network representation will be given in Section 15.

14. Conditional Independence

Define the *partners* of c_k to be the set D_k of propositional variables (other than c_k) that occur with c_k in K, for $k = 1, \ldots, n$. Let K_k be the set of constraints in K involving c_k and its preceding partners $c_j \in D_k$ such that $j < k$. Let $K_{\mathcal{L}_0} = \bigcup_{i=1}^m K_m$, the set of constraints in K only involving variables in $\mathcal{L}_0$. For A, B and C propositional variables or sets of propositional variables, write $I_p(A, B \mid C)$ if A and B are probabilistically independent conditional on C, according to p: C is said to *render A and B independent* or *screen off A* from B. We have the following useful properties:

THEOREM 14.1 (Conditional Independence Properties). *Suppose that K is a set of linear constraints and that p is determined by the entropic assignment ρ_e. Then*

1. *Let A be the set of variables which occur in $K_{\mathcal{L}_0}$ and let $B = \mathcal{L}_0 \backslash A$. If $K_{m+1}, \ldots, K_n$ are compatible with each probability function satisfying $K_{\mathcal{L}_0}$*

⋆ For specific and up-to-date details as to the computational properties of Bayesian networks, see the proceedings of the Uncertainty in Artificial Intelligence conferences, www.auai.org.

⋆⋆ See Williamson (2002b) for further analysis.

then for each variable $c \in B$ and for each state s of S such that $A \subseteq S \subseteq \mathcal{L}_0$, $p(c \mid s) = p_0(c \mid s)$. Consequently, if $I_{p_0}(c, F \mid E)$ for some disjoint subsets E and F of $\mathcal{L}_0$ such that $A \subseteq E \cup F$, then $I_p(c, F \mid E)$.

2. *If all the partners of the new propositional variables are in the old language, $D_k \subseteq \mathcal{L}_0$ for $k = m+1, \ldots, n$, then p is such that for $k = m+1, \ldots, n$, the partners D_k screen off c_k from all other variables, $I_p(c_k, \mathcal{L}\backslash(D_k \cup \{c_k\}) \mid D_k)$.*

Proof. First some terminology. Let $\mathcal{A}_k = \mathcal{A}\{c_1, \ldots, c_k\} = \{\pm c_1 \wedge \ldots \wedge \pm c_k\}$ be the atomic states involving $c_1, \ldots, c_k$. For $\alpha \in \mathcal{A}_k$ let c_i^α be the literal on c_i that is consistent with α. Thus we can write α as $c_1^\alpha \wedge \ldots \wedge c_k^\alpha$. Let α^i be $c_1^\alpha \wedge \ldots \wedge c_{i-1}^\alpha \wedge \neg c_i^\alpha \wedge c_{i+1}^\alpha \wedge \ldots \wedge c_k^\alpha$, the state in $\mathcal{A}_k$ that is identical to α except that it gives the opposite value to c_i. Recall that we write $\theta \sim \phi$ if sentences θ and ϕ are consistent.

The proof will be a constructive one, demonstrating how Lagrange multiplier methods can be used to determine p from the constraints. While this type of proof is far from brief, the Lagrange multiplier methods are important for practical calculations and the techniques presented here can be adapted for use in the algorithm of Section 15.

Part 1. This property just concerns the variables in $\mathcal{L}_0$ and is a feature of cross-entropy minimisation. We shall use Lagrange multipliers to minimise cross entropy subject to the constraints imposed by $K_{\mathcal{L}_0}$. Since the constraints in K are linear, $\mathbb{K}$ is convex and compact and the functions p that minimise cross entropy all agree on $\mathcal{L}_0$. The function $p|_{\mathcal{L}_0}$ will be shown to satisfy the conditional independence property. Since $K_{m+1}, \ldots, K_n$ are compatible with $p|_{\mathcal{L}_0}$, the function p subsequently produced by entropy maximisation extends $p|_{\mathcal{L}_0}$, and so will also satisfy the property.

Let $p_{0k}^\alpha = p_0(c_k^\alpha \mid c_1^\alpha \wedge \ldots \wedge c_{k-1}^\alpha)$ and $y_k^\alpha = p(c_k^\alpha \mid c_1^\alpha \wedge \ldots \wedge c_{k-1}^\alpha)$. Then to minimise cross entropy we must minimise

$$
\begin{aligned}
d_{\mathcal{L}_0}(p, p_0) &= \sum_{\beta \in \mathcal{A}\mathcal{L}_0} p(\beta) \log \frac{p(\beta)}{p_0(\beta)} \\
&= \sum_{\beta \in \mathcal{A}\mathcal{L}_0} \left(\prod_{i=1}^m y_i^\beta \right) \log \frac{\prod_{i=1}^m y_i^\beta}{\prod_{i=1}^m p_{0i}^\beta} \\
&= \sum_{\beta \in \mathcal{A}\mathcal{L}_0} \left(\prod_{i=1}^m y_i^\beta \right) \sum_{k=1}^m [\log y_k^\beta - \log p_{0k}^\beta] \\
&= \sum_{k=1}^m \sum_{\beta \in \mathcal{A}\mathcal{L}_0} \left(\prod_{i=1}^m y_i^\beta \right) \log \frac{y_k^\beta}{p_{0k}^\beta} \\
&= \sum_{k=1}^m \sum_{\alpha \in \mathcal{A}_k} \left(\prod_{i=1}^k y_i^\alpha \right) \log \frac{y_k^\alpha}{p_{0k}^\alpha}
\end{aligned}
$$

with respect to the parameters y_k^α. Now

$$\frac{\partial d_{\mathcal{L}_0}}{\partial y_k^\alpha} = \left(\prod_{i=1}^{k-1} y_i^\alpha\right)\left(1 + \log\frac{y_k^\alpha}{p_{0k}^\alpha} + \sum_{l=k+1}^{m}\sum_{\beta\in\mathcal{A}_l,\beta\sim\alpha}\left(\prod_{j=k+1}^{l} y_j^\beta\right)\left[\log\frac{y_l^\beta}{p_{0l}^\beta}\right]\right). \quad (1)$$

We have two types of constraint on cross entropy minimisation. Firstly there are those constraints χ in $K_{\mathcal{L}_0}$. Secondly, we also have constraints imposed by additivity of probability, namely $y_k^\alpha + y_k^{\alpha^k} = 1$ (i.e., $p(c_k^\alpha \mid c_1^\alpha \wedge \ldots \wedge c_{k-1}^\alpha) + p(\neg c_k^\alpha \mid c_1^\alpha \wedge \ldots \wedge c_{k-1}^\alpha) = 1$) for $k = 1, \ldots, m$ and each state $\alpha \in \mathcal{A}_k$. The Lagrange equation is

$$L = d_{\mathcal{L}_0} + \sum_{k=1}^{m}\sum_{\alpha\in\mathcal{A}_k}\mu_k^\alpha(y_k^\alpha + y_k^{\alpha^k} - 1) + \sum_{\chi\in K_{\mathcal{L}_0}}\lambda_\chi\chi,$$

where the μ_k^α are only dependent on k and $c_1^\alpha \wedge \ldots \wedge c_{k-1}^\alpha$ (i.e., $\mu_k^\alpha = \mu_k^{\alpha^k}$). Thus cross entropy is minimised if

$$\frac{\partial L}{\partial y_k^\alpha} = \frac{\partial d_{\mathcal{L}_0}}{\partial y_k^\alpha} + \mu_k^\alpha + \sum_{\chi\in K_{\mathcal{L}_0}}\lambda_\chi\frac{\partial\chi}{\partial y_k^\alpha} = 0.$$

We can cancel the μ_k^α term with the $\mu_k^{\alpha^k} = \mu_k^\alpha$ term in the equation involving state α^k to get:

$$\frac{\partial d_{\mathcal{L}_0}}{\partial y_k^\alpha} + \sum_{\chi\in K_{\mathcal{L}_0}}\lambda_\chi\frac{\partial\chi}{\partial y_k^\alpha} = \frac{\partial d_{\mathcal{L}_0}}{\partial y_k^{\alpha^k}} + \sum_{\chi\in K_{\mathcal{L}_0}}\lambda_\chi\frac{\partial\chi}{\partial y_k^{\alpha^k}} \quad (2)$$

for each $k = 1, \ldots, m$ and $\alpha \in \mathcal{A}_k$. Equation (2) together with the constraints χ determine the values of the parameters y_k^α.

Without loss of generality we shall suppose that the variables in $\mathcal{L}_0$ are ordered so that the variables that occur in $K_{\mathcal{L}_0}$ come first in the ordering: $\mathcal{L}_0 = \{c_1, \ldots, c_t, c_{t+1}, \ldots, c_m\}$ where $c_1, \ldots, c_t$ are the variables that occur in $K_{\mathcal{L}_0}$. By assumption each constraint χ in $K_{\mathcal{L}_0}$ is linear and can then be written

$$\sum_{i=1}^{r} a_i\, p(\theta_i) - b = 0$$

$$\Leftrightarrow \sum_{i=1}^{r} a_i \sum_{\alpha\in\mathcal{A}_t,\alpha\sim\theta_i} p(\alpha) - b = 0$$

$$\Leftrightarrow \sum_{i=1}^{r} a_i \sum_{\alpha\in\mathcal{A}_t,\alpha\sim\theta_i}\prod_{j=1}^{t} y_j^\alpha - b = 0.$$

Then for $k = 1, \ldots, t$,

$$\frac{\partial \chi}{\partial y_k^\alpha} = \sum_{\beta \in \mathcal{A}_t, \beta \sim \alpha} A^\beta \prod_{i=1, i \neq k}^{t} y_i^\beta,$$

where the constant $A^\beta = \sum_{i:\beta \sim \theta_i} a_i$ ($= 0$ if $\beta \not\sim \theta_i$ for any i).

Note especially that for $k = t + 1, \ldots, m$,

$$\frac{\partial \chi}{\partial y_k^\alpha} = 0.$$

We shall now show that for $k = t + 1, \ldots, m$, $y_k^\alpha = p_{0k}^\alpha$, by induction from $k = m$ down to $k = t + 1$.

For $k = m$ Equation (1) reduces to

$$\frac{\partial d_{\mathcal{L}_0}}{\partial y_k^\alpha} = \left(\prod_{i=1}^{k-1} y_i^\alpha \right) \left(1 + \log \frac{y_k^\alpha}{p_{0k}^\alpha} \right)$$

and so Equation (2) reduces to

$$\log \frac{y_k^\alpha}{p_{0k}^\alpha} = \log \frac{y_k^{\alpha^k}}{p_{0k}^{\alpha^k}}$$

which in turn implies that $y_k^\alpha = p_{0k}^\alpha$ or $p_{0k}^\alpha = 0$. However, if $p_{0k}^\alpha = 0$ then by open-mindedness $y_k^\alpha = 0$, so $y_k^\alpha = p_{0k}^\alpha$ in either eventuality.

For $t < k < m$ we get a similar reduction. By the induction hypothesis $y_l^\beta = p_{0l}^\beta$ for $l = k + 1, \ldots, m$, so terms in Equation (1) involving $\log(y_l^\beta / p_{0l}^\beta)$ vanish and again

$$\frac{\partial d_{\mathcal{L}_0}}{\partial y_k^\alpha} = \left(\prod_{i=1}^{k-1} y_i^\alpha \right) \left(1 + \log \frac{y_k^\alpha}{p_{0k}^\alpha} \right).$$

Thus Equation (2) implies again that $y_k^\alpha = p_{0k}^\alpha$.

Now suppose we are given any $c \in B$ and set $S \supseteq A$. If $c \in S$ then for any state s of S, $p(c \mid s) = p_0(c \mid s) \in \{0, 1\}$ by the axioms of probability. If $c \notin S$, order the variables so that $A = \{c_1, \ldots, c_t\}$ (as above), $S = \{c_1, \ldots, c_{k-1}\}$ and $c = c_k$. Then the fact that $y_k^\alpha = p_{0k}^\alpha$ for each α means that $p(c \mid s) = p_0(c \mid s)$ for each state s of S.

Next to prove the conditional independence property. Let E and F be subsets of $\mathcal{L}_0$ such that $S = E \cup F \supseteq A$, and suppose that $I_{p_0}(c, F \mid E)$. This occurs if and only if $p_0(c \mid s)$ stays constant as state s of S varies over F. We know now that $p(c \mid s) = p_0(c \mid s)$ for each such s, and so $I_p(c, F \mid E)$, as required.

Part 2. Since the constraints in K are linear, $\mathbb{K}$ is convex and compact and the restriction of p to $\mathcal{L}_0$ is fixed uniquely by cross entropy minimisation. It will not

matter here what function this restriction is: this conditional independence property is a result of the next stage in the process, entropy maximisation. We shall use Lagrange multipliers to determine the function p which maximises entropy, just as in Part 1.

Now p is found by maximising

$$H_{c_{m+1},\ldots,c_n|c_1,\ldots,c_m}(p) = \sum_{k=m+1}^{n} H_{c_k|c_1,\ldots,c_{k-1}}(p)$$

$$= -\sum_{k=m+1}^{n} \sum_{\alpha\in\mathcal{A}_k} p^\alpha \left(\prod_{i=m+1}^{k-1} y_i^\alpha\right)[y_k^\alpha \log y_k^\alpha]$$

with respect to parameters $y_k^\alpha = p(c_k^\alpha \mid c_1^\alpha \wedge \ldots \wedge c_{k-1}^\alpha)$, for each $\alpha \in \mathcal{A}_k$, and where $p^\alpha = p(c_1^\alpha \wedge \ldots \wedge c_m^\alpha)$, a constant, fixed by having minimised cross entropy.

Again we have two types of constraint on entropy maximisation: the constraints χ in K, and the additivity constraints imposed by additivity $y_k^\alpha + y_k^{\alpha^k} = 1$ for $k = m+1, \ldots, n$ and each state $\alpha \in \mathcal{A}_k$. Therefore the Lagrange equation is:

$$L = \sum_{k=m+1}^{n} \sum_{\alpha\in\mathcal{A}_k} \left(p^\alpha \left(\prod_{i=m+1}^{k-1} y_i^\alpha\right)[y_k^\alpha \log y_k^\alpha] + \mu_k^\alpha (y_k^\alpha + y_k^{\alpha^k} - 1)\right)$$

$$+ \sum_{\chi\in K} \lambda_\chi \chi. \tag{3}$$

(Here we adopt the usual convention that $0 \log 0 = 0$.)

Recall that we have divided K into $K_{\mathcal{L}_0}, K_{m+1}, \ldots, K_n$, where $K_{\mathcal{L}_0}$ contains constraints only involving propositional variables in $\mathcal{L}_0$ and K_k contains constraints only involving c_k and its partners, for $k = m+1, \ldots, n$. Note that by assumption, $D_k \subseteq \mathcal{L}_0$, and so c_k only appears in constraints in K_k, for $k = m+1, \ldots, n$. Let α run through the states of $\{c_1, \ldots, c_k\}$. By assumption χ each constraint in K_k is linear and can then be written

$$\sum_{i=1}^{r} a_i\, p(\theta_i) - b = 0$$

$$\Leftrightarrow \sum_{i=1}^{r} a_i \sum_{\alpha\in\mathcal{A}_k, \alpha\sim\theta_i} p(\alpha) - b = 0$$

$$\Leftrightarrow \sum_{i=1}^{r} a_i \sum_{\alpha\in\mathcal{A}_k, \alpha\sim\theta_i} p^\alpha \prod_{j=m+1}^{k} y_j^\alpha - b = 0$$

and each constraint χ in $K_{\mathcal{L}_0}$ is likewise of the form

$$\sum_{i=1}^{r} a_i \sum_{\alpha\in\mathcal{A}_{\mathcal{L}_0}, \alpha\sim\theta_i} p^\alpha - b = 0.$$

Then for $k = m+1, \ldots, n$ and $\chi \in K_l$ where $l < k$,

$$\frac{\partial \chi}{\partial y_k^\alpha} = 0.$$

For $\chi \in K_k$,

$$\frac{\partial \chi}{\partial y_k^\alpha} = A^\alpha p^\alpha \prod_{i=m+1}^{k-1} y_i^\alpha,$$

where again the constant $A^\alpha = \sum_{i:\alpha \sim \theta_i} a_i$ ($= 0$ if $\alpha \not\sim \theta_i$ for any i). Finally for $\chi \in K_l$ where $l > k$,

$$\begin{aligned}\frac{\partial \chi}{\partial y_k^\alpha} &= \sum_{\beta \in \mathcal{A}_l, \beta \sim \alpha} A^\beta p^\beta \prod_{i=m+1, i \neq k}^{l} y_i^\beta \\ &= p^\alpha \prod_{i=m+1}^{k-1} y_i^\alpha \sum_{\beta \in \mathcal{A}_l, \beta \sim \alpha} A^\beta \prod_{i=k+1}^{l} y_i^\beta.\end{aligned}$$

Notice that if β and β' agree on the propositional variables which occur in $\chi \in K_l$ (and hence if β and β' agree on D_l) then $A^\beta = A^{\beta'}$.

Thus entropy is maximised if

$$\begin{aligned}\frac{\partial L}{\partial y_k^\alpha} &= \mu_k^\alpha + p^\alpha \left(\prod_{i=m+1}^{k-1} y_i^\alpha \right) \\ &\quad \times \left(1 + \log y_k^\alpha + \sum_{\chi \in K_k} \lambda_\chi A_\chi^\alpha + \sum_{l=k+1}^{n} \sum_{\beta \in \mathcal{A}_l, \beta \sim \alpha} \left(\prod_{j=k+1}^{l} y_j^\beta \right) \right. \\ &\quad \left. \times \left[\log y_l^\beta + \sum_{\chi \in K_l} \lambda_\chi A_\chi^\beta \right] \right) \\ &= 0.\end{aligned}$$

Since $\mu_k^\alpha = \mu_k^{\alpha^k}$ these terms cancel when we form the equation $\partial L / \partial y_k^\alpha = \partial L / \partial y_k^{\alpha^k}$. Since $p^\alpha = p^{\alpha^k}$ and $y_i^\alpha = y_i^{\alpha^k}$ for $k = m+1, \ldots, n$, $i = m+1, \ldots, k-1$, we get

$$\begin{aligned}&\log \frac{y_k^\alpha}{y_k^{\alpha^k}} + \sum_{\chi \in K_k} \lambda_\chi (A_\chi^\alpha - A_\chi^{\alpha^k}) \\ &\quad + \sum_{l=k+1}^{n} \sum_{\beta \in \mathcal{A}_l, \beta \sim \alpha} \left(\prod_{j=k+1}^{l} y_j^\beta \right) \left[\log y_l^\beta + \sum_{\chi \in K_l} \lambda_\chi A_\chi^\beta \right]\end{aligned}$$

$$= \sum_{l=k+1}^{n} \sum_{\beta \in \mathcal{A}_l, \beta \sim \alpha} \left(\prod_{j=k+1}^{l} y_j^{\beta^k} \right) \left[\log y_l^{\beta^k} + \sum_{\chi \in K_l} \lambda_\chi A_\chi^{\beta^k} \right].$$

We can now show that the independence relation $I(c_k, \mathcal{L} \backslash (D_k \cup \{c_k\}) \mid D_k)$ holds by induction on k, from n down to $m+1$.

For $k = n$ the above equation reduces to

$$\log \frac{y_k^\alpha}{y_k^{\alpha^k}} + \sum_{\chi \in K_k} \lambda_\chi (A_\chi^\alpha - A_\chi^{\alpha^k}) = 0.$$

By eliminating the Lagrange multipliers we find that

$$\log \frac{y_k^\alpha}{y_k^{\alpha^k}} = \log \frac{y_k^{\alpha'}}{y_k^{\alpha'^k}}$$

for each pair of states α, α' consistent with the same state of $D_k \cup \{c_k\}$. This implies that $y_k^\alpha = y_k^{\alpha'}$, the required conditional independence relationship.

For $m < k < n$ we get a similar reduction. By the induction hypothesis $y_j^{\beta^k} = y_j^\beta$ for each $j = k+1, \ldots, n$, since $c_k \notin D_j$. Further, for $\chi \in K_j$, $j > k$, we have that $A_\chi^{\beta^k} = A_\chi^\beta$, since c_k does not occur in χ. Then,

$$\sum_{l=k+1}^{n} \sum_{\beta \in \mathcal{A}_l, \beta \sim \alpha} \left(\prod_{j=k+1}^{l} y_j^{\beta} \right) \left[\log y_l^{\beta} + \sum_{\chi \in K_l} \lambda_\chi A_\chi^{\beta} \right]$$

$$= \sum_{l=k+1}^{n} \sum_{\beta \in \mathcal{A}_l, \beta \sim \alpha} \left(\prod_{j=k+1}^{l} y_j^{\beta^k} \right) \left[\log y_l^{\beta^k} + \sum_{\chi \in K_l} \lambda_\chi A_\chi^{\beta^k} \right].$$

Hence, just as in the case $k = n$,

$$\log \frac{y_k^\alpha}{y_k^{\alpha^k}} + \sum_{\chi \in K_k} \lambda_\chi (A_\chi^\alpha - A_\chi^{\alpha^k}) = 0,$$

$$\log \frac{y_k^\alpha}{y_k^{\alpha^k}} = \log \frac{y_k^{\alpha'}}{y_k^{\alpha'^k}}$$

and so $y_k^\alpha = y_k^{\alpha'}$ for α, α' consistent with the same state of $D_k \cup \{c_k\}$, as required. □

Note in particular that if there is only one new variable ($n = m+1$), and the constraints in K are linear, then the partners of the new variable must all be in $\mathcal{L}_0$, so the entropic assignment will render the new variable probabilistically independent of all variables except its partners, conditional on its partners.

Note also that the set of partners of c_k depends on the way that the constraints in K are formulated. For example, K may consist just of the constraint $p(c_n) = p(c_0)$, in which case $D_n = \{c_0\}$. But the same constraint may be written $p(c_n \wedge c_1) + p(c_n \wedge \neg c_1) = p(c_0)$, in which case $D_n = \{c_0, c_1\}$. The conditional independence property implies in the first case that $I(c_n, c_j \mid c_0)$ for $j = 1, \ldots, n-1$, and in the second case that $I(c_n, c_j \mid c_0, c_1)$ for $j = 2, \ldots, n-1$ (assuming that $m \geq 1$). But these conclusions are quite consistent since the latter independence relation follows from the former by the axioms of probability (Pearl, 1988: §3.1.4).

In Section 15 we shall show how the conditional independence properties enable us to construct a Bayesian network representation of p. Finally in Section 16 we shall see that the condition in Theorem 14.1 requiring constraints to be linear is not, in fact, necessary.

15. Bayesian Network Representation

We have seen how the entropic assignment preserves certain conditional independencies in the initial probability function p_0, and that it can also render new variables conditionally independent. Bayesian networks represent probability functions efficiently by exploiting conditional independencies. Thus it seems plausible that the entropic assignment might allow efficient belief change in the Bayesian network framework. In this section we shall see that this is so, first through a series of examples, and then by providing an algorithm for constructing a suitable Bayesian network representation.

The key to efficient belief change lies in replacing the expensive operations of minimising cross entropy and maximising entropy over the whole language by local optimisations. The key advantages of such representations are computational. Suppose p_0 is represented by a Bayesian network, the transitional knowledge K is linear and compatible with p_0, and all partners of new variables are in the old language. We can then apply the entropic assignment simply by taking the old network, adding the new nodes, adding arrows to each new node from each of its partners, and adding new probability specifiers $p(c_k \mid D_k^\alpha)$, for $k = m+1, \ldots, n$, by maximising entropy locally on $D_k \cup \{c_k\}$. If the size of each D_k is small relative to n then the space taken up by such a representation, the time taken to maximise entropy to find the new specifiers, and the time taken to calculate required probabilities from the new Bayesian network may all be reduced to practical levels.

In our examples, $\mathcal{L}_0 = \{c_1, c_2, c_3\}$ and $\mathcal{L}_+ = \{c_4, c_5\}$. p_0 is represented by a Bayesian network with graph as in Figure 1 and probability specification $p_0(c_1) = 0.2$, $p_0(c_2 \mid c_1) = 0.9$, $p_0(c_2 \mid \neg c_1) = 0.4$, $p_0(c_3 \mid c_1) = 0.8$, $p_0(c_3 \mid \neg c_1) = 0.1$.

Example 15.1. Suppose the transitional knowledge is of the form $K = \{p(c_4 \mid c_1) = 0.6\}$. K is compatible with p_0 on $\mathcal{L}_0$, so $p|_{\mathcal{L}_0} = p_0$ and the Bayesian network representing p will extend the initial network representing p_0. Theorem 14.1.2 implies that $I_p(c_4, \{c_2, c_3, c_5\} \mid c_1)$ and $I_p(c_5, \{c_1, c_2, c_3, c_4\})$. These conditional

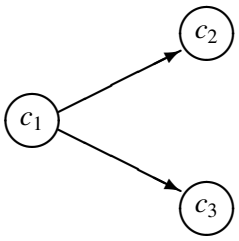

Figure 1. Graph in the Bayesian network representation of p_0.

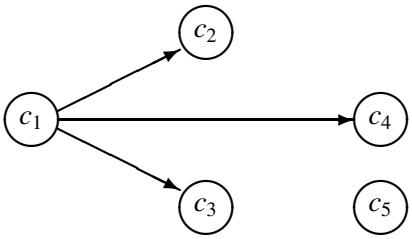

Figure 2. Graph in the Bayesian network representation of p.

independencies can be used to construct a graph in a Bayesian network representation of p, as in Figure 2. All that remains is to add the probability specifiers of c_4 conditional on c_1, and of c_5. These are found by maximising entropy locally on the new nodes. Equation (3) becomes

$$L = \sum_{\alpha \in \mathcal{A}\{c_1, c_4\}} p^\alpha y_4^\alpha \log y_4^\alpha + \mu_4^\alpha (y_4^\alpha + y_4^{\alpha^4} - 1) + \lambda (y_4^{c_1 \wedge c_4} - 0.6)$$
$$+ \sum_{\alpha \in \mathcal{A}\{c_5\}} p^\alpha y_5^\alpha \log y_5^\alpha + \mu_5^\alpha (y_5^\alpha + y_5^{\alpha^5} - 1).$$

Now $y_4^{c_1 \wedge c_4}$, $y_4^{c_1 \wedge \neg c_4}$ are fully constrained, while the pairs $y_4^{\neg c_1 \wedge c_4}$, $y_4^{\neg c_1 \wedge \neg c_4}$ and $y_5^{c_5}$, $y_5^{\neg c_5}$ are only constrained by additivity. It is straightforward to see that in such a situation entropy is maximised when

$$y_4^{c_1 \wedge c_4} = p(c_4 \mid c_1) = 0.6, \quad y_4^{c_1 \wedge \neg c_4} = p(\neg c_4 \mid c_1) = 0.4,$$

$$y_4^{\neg c_1 \wedge c_4} = p(c_4 \mid \neg c_1) = 0.5 = y_4^{\neg c_1 \wedge \neg c_4} = p(\neg c_4 \mid \neg c_1),$$

$$y_5^{c_5} = p(c_5) = 0.5 = y_5^{\neg c_5} = p(\neg c_5).$$

These values complete the probability specification of the Bayesian network, which now fully determines p.

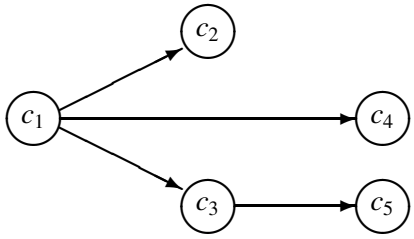

Figure 3. Graph in the Bayesian network representation of p.

Example 15.2. Suppose the transitional knowledge is now $K = \{p(c_4 \mid c_1) = 0.6,\ p(c_5 \wedge \neg c_3) = 2p(c_5 \wedge c_3)\}$. K is still compatible with p_0 on $\mathcal{L}_0$, so we again merely extend the Bayesian network representing p_0. Theorem 14.1.2 implies that $I_p(c_4, \{c_2, c_3, c_5\} \mid c_1)$ and $I_p(c_5, \{c_1, c_2, c_4\} \mid c_3)$. These conditional independencies are represented by Figure 3. To determine the new probability specifiers we maximise entropy locally on c_4, c_1 and c_5, c_3. Equation (3) becomes

$$\begin{aligned} L \ = \ & \sum_{\alpha \in \mathcal{A}\{c_1, c_4\}} p^\alpha y_4^\alpha \log y_4^\alpha + \mu_4^\alpha (y_4^\alpha + y_4^{\alpha^4} - 1) + \lambda_1 (y_4^{c_1 \wedge c_4} - 0.6) \\ & + \sum_{\alpha \in \mathcal{A}\{c_3, c_5\}} p^\alpha y_5^\alpha \log y_5^\alpha + \mu_5^\alpha (y_5^\alpha + y_5^{\alpha^5} - 1) \\ & + \lambda_2 (p^{\neg c_3 \wedge c_5} y_5^{\neg c_3 \wedge c_5} - 2p^{c_3 \wedge c_5} y_5^{c_3 \wedge c_5}). \end{aligned}$$

As above, $y_4^{c_1 \wedge c_4} = 0.6$, $y_4^{c_1 \wedge \neg c_4} = 0.4$, so we just need to focus on the parameters y_5^α. We thus have

$$\frac{\partial L}{\partial y_5^\alpha} = p^\alpha (1 + \log y_5^\alpha) + \mu_5^\alpha + k\lambda_2,$$

where

$$k = \begin{cases} p^{\neg c_3 \wedge c_5}: & \alpha = \neg c_3 \wedge c_5 \\ -2p^{c_3 \wedge c_5}: & \alpha = c_3 \wedge c_5 \\ 0: & \text{otherwise} \end{cases}$$

Noting that $p^\alpha = p^{\alpha^\kappa}$ we eliminate $\mu_5^\alpha = \mu_5^{\alpha^k}$ to obtain the solution

$$y_5^{c_3 \wedge c_5} = 0.66, \quad y_5^{c_3 \wedge \neg c_5} = 0.34, \quad y_5^{\neg c_3 \wedge c_5} = 0.42, \quad y_5^{\neg c_3 \wedge \neg c_5} = 0.58.$$

Example 15.3. Now the transitional knowledge $K = \{p(c_4 \mid c_1) = 0.6,\ p(c_5 \wedge \neg c_3) = 2p(c_5 \wedge c_3),\ p(c_1) = 0.7\}$. This transitional knowledge is no longer compatible with p_0 on $\mathcal{L}_0$, and so minimising cross-entropy is non-trivial. Note however that $K_4 = \{p(c_4 \mid c_1) = 0.6\}$ and $K_5 = \{p(c_5 \wedge \neg c_3) = 2p(c_5 \wedge c_3)\}$ are compatible with each function satisfying $K_{\mathcal{L}_0} = \{p(c_1) = 0.7\}$, and so Theorem 14.1.1 applies. Thus we know that $I_p(c_2, c_3 \mid c_1)$, which ensures that Figure

3 remains the appropriate structure, and we know that $p(c_2 \mid \pm c_1) = p_0(c_2 \mid \pm c_1)$ and $p(c_3 \mid \pm c_1) = p_0(c_3 \mid \pm c_1)$, and so to determine the Bayesian network representing p on $\mathcal{L}_0$, all we need do is determine the specifier $p(c_1)$. This is fixed by $K_{\mathcal{L}_0} = \{p(c_1) = 0.7\}$. Next we maximise entropy. This is done in the same way as the last example, except now $p(c_1) = 0.7$ so

$$y_5^{c_3 \wedge c_5} = 0.23, \quad y_5^{c_3 \wedge \neg c_5} = 0.77, \quad y_5^{\neg c_3 \wedge c_5} = 0.65, \quad y_5^{\neg c_3 \wedge \neg c_5} = 0.35.$$

Example 15.4. In this example the transitional knowledge $K = \{p(c_4 \mid c_1) = 0.6, p(c_5 \wedge \neg c_3) = 2p(c_5 \wedge c_3), p(c_1 \wedge c_2) = 0.4\}$. Again, the transitional knowledge is incompatible with p_0 on $\mathcal{L}_0$, and we have work to do to minimise cross-entropy. We still have that $I_p(c_2, c_3 \mid c_1)$, and so Figure 3 represents the conditional independencies that p satisfies on $\mathcal{L}_0$. Moreover, Theorem 14.1.1 implies that $p(c_3 \mid \pm c_1) = p(c_3 \mid \pm c_1 \wedge \pm c_2) = p_0(c_3 \mid \pm c_1 \wedge \pm c_2) = p_0(c_3 \mid \pm c_1)$. Hence we need to minimise cross entropy locally on c_1 and c_2. Writing the constraint $y_2^{c_1 \wedge c_2} y_1^{c_1} = 0.4$, Equation (2) gives us the equations

$$\log \frac{y_1^{c_1}}{p_{01}^{c_1}} + y_2^{c_1 \wedge \neg c_2} \log \frac{y_2^{c_1 \wedge \neg c_2}}{p_{02}^{c_1 \wedge \neg c_2}} + y_2^{c_1 \wedge c_2} \log \frac{y_2^{c_1 \wedge c_2}}{p_{02}^{c_1 \wedge c_2}} + \lambda y_2^{c_1 \wedge c_2}$$
$$= \log \frac{y_1^{\neg c_1}}{p_{01}^{\neg c_1}} + y_2^{\neg c_1 \wedge \neg c_2} \log \frac{y_2^{\neg c_1 \wedge \neg c_2}}{p_{02}^{\neg c_1 \wedge \neg c_2}} + y_2^{\neg c_1 \wedge c_2} \log \frac{y_2^{\neg c_1 \wedge c_2}}{p_{02}^{\neg c_1 \wedge c_2}},$$

$$\log \frac{y_2^{\neg c_1 \wedge \neg c_2}}{p_{02}^{\neg c_1 \wedge \neg c_2}} = \log \frac{y_2^{\neg c_1 \wedge c_2}}{p_{02}^{\neg c_1 \wedge c_2}},$$

$$\log \frac{y_2^{c_1 \wedge \neg c_2}}{p_{02}^{c_1 \wedge \neg c_2}} = \log \frac{y_2^{c_1 \wedge c_2}}{p_{02}^{c_1 \wedge c_2}} + \lambda.$$

By eliminating λ and using the transitional knowledge constraint and additivity constraints we see that:

$$y_1^{\neg c_1} = 0.50 = y_1^{c_1},$$

$$y_2^{\neg c_1 \wedge \neg c_2} = p_{02}^{\neg c_1 \wedge \neg c_2} = 0.6, \quad y_2^{\neg c_1 \wedge c_2} = p_{02}^{\neg c_1 \wedge c_2} = 0.4,$$

$$y_2^{c_1 \wedge \neg c_2} = 0.20, \quad y_2^{\neg c_1 \wedge \neg c_2} = 0.80.$$

These we add to the probability specification of the Bayesian network over $\mathcal{L}_0$. We then conclude by maximising entropy as in the previous example.

The approach taken in these examples can be generalised as follows.

The *Network Assignment Algorithm*:

Input $\mathfrak{B}_0 = (G_0, S_0)$, a Bayesian network representing p_0 over $\mathcal{L}_0$; K, the transitional knowledge. Assume that the conditions of Theorem 14.1 are satisfied.*

- If K is compatible with p_0 over $\mathcal{L}_0$, set $G = G_0$, $S = S_0$.
- Else:
  - set $G = G_0$,
  - add arrows to ensure each pair of variables in $K_{\mathcal{L}_0}$ is connected by an arrow,
  - for each variable c_k that is in $K_{\mathcal{L}_0}$ or an ancestor of a variable in $K_{\mathcal{L}_0}$, determine the specifiers corresponding to c_k: these are the parameters y_k^α, for states α of c_k and its parents, that solve the system of equations given by the constraints in $K_{\mathcal{L}_0}$, the additivity constraints $y_k^\alpha + y_k^{\alpha^k} = 1$, and the minimum cross entropy constraints as in Equation (2).
  - for each other variable, add to S the specifiers of that variable given in S_0.
- For each new variable c_k, $k = m+1, \ldots, n$,
  - add c_k as a node in G,
  - add an arrow from each partner of c_k to c_k in G,
  - determine the corresponding probability specifiers: these are the parameters y_k^α, $\alpha \in \mathcal{A}(D_k \cup \{c_k\})$, which solve the system of equations given by the constraints in K_k, the additivity constraints $y_k^\alpha + y_k^{\alpha^k} = 1$, and the maximum entropy constraints $\log y_k^\alpha / y_k^{\alpha^k} + \sum_{\chi \in K_k} \lambda_\chi (A_\chi^\alpha - A_\chi^{\alpha^k}) = 0$.

Output $\mathfrak{B} = (G, S)$, a Bayesian network representing p over $\mathcal{L}$.

THEOREM 15.5. *The network assignment algorithm is correct: it determines a Bayesian network representing p from a network representing p_0 and suitable transitional knowledge.*

Proof. Let $Anc(X)$, $Des(X)$, $Par(X)$ and $Chi(X)$ be respectively the ancestors, descendants, parents and children of variables in X. Let $Nid(X)$ be the set of variables which are not in the immediate family or descendants of X, i.e., all variables except for those in X, the parents of variables in X, and the descendants of variables in X. These relationships will be assumed to be determined by G_0, unless indicated otherwise, and the abbreviations Anc_k, Des_k etc. refer to $Anc(c_k)$, $Des(c_k)$ etc.. We shall go through the algorithm step by step.

* We shall see in Section 16 that the assumption of linear constraints may be relaxed.

If K is compatible with p_0 over $\mathcal{L}_0$, then p extends p_0 (Proposition 11.1), and so G extends G_0 and S extends S_0.

Otherwise cross entropy minimisation is non-trivial. Let A be the set of variables occurring in $K_{\mathcal{L}_0}$. The constraints $K_{\mathcal{L}_0}$ on $\mathcal{L}_0$ may render all the variables in A dependent, and so these variables need to be connected in G. All other connections remain the same for the following reasons. If $c_k \in \mathcal{L}_0 \backslash (A \cup Anc(A))$ we have that $I_{p_0}(c_k, Nid_k \mid Par_k)$ and $A \subseteq Nid_k \cup Par_k$ so by Theorem 14.1.1, $I_p(c_k, Nid_k \mid Par_k)$ and for each state α of $\mathcal{L}_0$, $p(c_k^\alpha \mid Par_k^\alpha) = p(c_k^\alpha \mid Nid_k^\alpha \wedge Par_k^\alpha) = p_0(c_k^\alpha \mid Nid_k^\alpha \wedge Par_k^\alpha) = p_0(c_k^\alpha \mid Par_k^\alpha)$. Thus for those variables not in A or ancestors of A, the graphical connections and specifiers remain unchanged. To show that the connections amongst the ancestors of A and those linking the ancestors of A to A remain unchanged it is enough (see Cowell et al., 1999: theorem 5.14). to show that for each variable $c_k \in Anc(A)$ the Markov blanket $Mb_k = Par_k \cup Chi_k \cup Par(Chi_k)$ in G_0 continues to screen c_k off from the other variables. To this end we have that $I_{p_0}(c_k, \mathcal{L}_0 \backslash (Mb_k \cup c_k) \mid Mb_k)$ and $A \subseteq \mathcal{L}_0 \backslash c_k$ so by Theorem 14.1.1, $I_p(c_k, \mathcal{L}_0 \backslash (Mb_k \cup c_k) \mid Mb_k)$.

The next step is to determine the specifiers corresponding to variables in A and their ancestors. The specifiers of the other variables in $\mathcal{L}_0$ are, as demonstrated above, the same as those in $\mathfrak{B}_0$.

We move on to the entropy maximisation phase. We add an arrow to each new variable c_k from its partners because we know by Theorem 14.1.2 that the partners screen c_k off from other variables, and these other variables constitute its non-descendants. Thus $I_p(c_k, Nid_k \mid Par_k)$. The final step is to determine the corresponding probability specifiers. □

16. Freedom from Constraints

In this section I shall extend Theorem 14.1 to the case where constraints may be non-linear. The proof of Theorem 14.1 uses Lagrange multipliers to construct the probability function determined by the entropic assignment, and is useful in itself for explicitly showing how cross entropy can be minimised and entropy maximised. In this section we will pursue a less constructive approach which allows us to focus on the nature of constraints and how they interact.

The key concept of this section is that of *freedom*: roughly speaking a set A of variables is free in $\mathbb{K}$ if their probabilities are unconstrained in $\mathbb{K}$. There are in fact several ways this notion can be explicated, depending on how we interpret "their probabilities" in the previous sentence. If the marginal probability distribution of A is unconstrained we say that A is *marginally* free. If the probabilities that determine the relationships between variables in A and those in $B = \mathcal{L} \backslash A$ are unconstrained, A is *relatively* free. If both the distribution of A and A's relationships with B are unconstrained, A is *jointly* free. In each case "unconstrained" in $\mathbb{K}$ means that if we take a probability function p in $\mathbb{K}$ and hold fixed the marginal probability

distribution over B, we can vary other probabilities arbitrarily and still remain within $\mathbb{K}$.

In what follows the set A is always assumed to be a subset of $\mathcal{L}$, $B = \mathcal{L}\backslash A$, $n = |\mathcal{L}|$, for $\alpha \in \mathcal{A}A$ and $\beta \in \mathcal{A}B$ concatenations $\alpha\beta$ are to be read as conjunctions $\alpha \wedge \beta$ (so that the concatenations $\alpha\beta$ are the atomic states of $\mathcal{L}$), and elements of $[0, 1]^{2^n}$ are indexed by the atomic states of $\mathcal{L}$.

DEFINITION 16.1. *$A \subseteq \mathcal{L}$ is* marginally free *in $\mathbb{K}$ if*

- *given any $p \in \mathbb{K}$ and $x \in [0, 1]^{2^{|A|}}$ such that $\sum_{\alpha \in \mathcal{A}A} x_\alpha = 1$,*
- *there exists some $q \in \mathbb{K}$ such that $q(\beta) = p(\beta)$ for each $\beta \in \mathcal{A}B$, and $q(\alpha) = x_\alpha$ for each $\alpha \in \mathcal{A}A$.*

DEFINITION 16.2. *A is* relatively free *in $\mathbb{K}$ if*

- *given any $p \in \mathbb{K}$ and $x \in [0, 1]^{2^n}$ such that $\sum_\beta x_{\alpha\beta} = p(\alpha)$ and $\sum_\alpha x_{\alpha\beta} = p(\beta)$ for each α, β,*
- *there exists some $q \in \mathbb{K}$ such that $q(\alpha\beta) = x_{\alpha\beta}$ for each α, β.*

DEFINITION 16.3. *A is* jointly free *in $\mathbb{K}$ if*

- *given any $p \in \mathbb{K}$ and $x \in [0, 1]^{2^n}$ such that $\sum_\alpha x_{\alpha\beta} = p(\beta)$ for each β,*
- *there exists some $q \in \mathbb{K}$ such that $q(\alpha\beta) = x_{\alpha\beta}$ for each α, β.*

These definitions imply:

PROPOSITION 16.4.

1. *A is jointly free in $\mathbb{K}$ if and only if it is marginally free in $\mathbb{K}$ and relatively free in $\mathbb{K}$.*
2. *If A is marginally free in $\mathbb{K}$ then $\mathbb{K}$ is compatible on A. (The converse is not true.)*
3. *If A is jointly free in $\mathbb{K}$ and $C \subseteq A$ then C is jointly free in $\mathbb{K}$.*
4. *If A is jointly (or marginally or relatively) free in $\mathbb{K}$ with respect to language $\mathcal{L}$, then $A \cap \mathcal{L}_0$ is jointly (respectively, marginally or relatively) free in $\mathbb{K}_0$ with respect to $\mathcal{L}_0$.*
5. *A is jointly free in $\mathbb{K}$ if and only if,*

   - *given any $y \in [0, 1]^{2^n}$ such that $\sum_\alpha y_{\alpha\beta} = 1$ (for all β),*
   - *there is some $p \in \mathbb{K}$ such that $p(\beta) > 0 \Rightarrow p(\alpha \mid \beta) = y_{\alpha\beta}$ (for all α, β).*[⋆]

[⋆] In Section 10 I took the line that if $p(\phi) = 0$ then $p(\theta \mid \phi)$ is unconstrained (rather than undefined). This can be explicated thus: given $p \in \mathbb{K}$, any $\phi \in \mathcal{SL}$ such that $p(\phi) = 0$, any $\theta \in \mathcal{SL}$,

One generalisation of the concept of relative freedom will be important in what follows.

DEFINITION 16.5. *Subsets* $A_1, \ldots, A_k$ *of* $\mathcal{L}$ *are* relatively free *in* $\mathbb{K}$ *if (setting* $A = \bigcup_{i=1}^{k} A_i$*)*

- *given any* $p \in \mathbb{K}$ *and* $x \in [0,1]^{2^{|A|}}$ *such that for all* $i = 1, \ldots, k$ *and* $\alpha_i \in \mathcal{A}A_i$, $\sum_{\alpha \sim \alpha_i} x_\alpha = p(\alpha_i)$,
- ▶ *there exists some* $q \in \mathbb{K}$ *such that* $q(\alpha) = x_\alpha$ *for all* α.

This definition gives:

PROPOSITION 16.6.

1. *A is relatively free in* $\mathbb{K}$ *(according to Definition 16.2) iff A and* $\mathcal{L} \backslash A$ *are relatively free in* $\mathbb{K}$ *(according to Definition 16.5).*
2. *Suppose* $A_1, \ldots, A_k$ *are relatively free in* $\mathbb{K}$ *with respect to* $\mathcal{L}$*, and are pairwise disjoint in* $\mathcal{L}_+$.[⋆] *Let* $B_i = A_i \cap \mathcal{L}_0$*, for* $i = 1, \ldots, k$*. Then* $B_1, \ldots, B_k$ *are relatively free in* $\mathbb{K}_0$ *with respect to* $\mathcal{L}_0$.

Now suppose each constraint χ is of the form $f(p(\alpha_1), \ldots, p(\alpha_{2^l})) = 0$, where the $\alpha_1, \ldots, \alpha_{2^l}$ are the atomic states of set $A = \{c_{i_1}, \ldots, c_{i_l}\}$. Thus χ is a constraint on the marginal distribution of A.

PROPOSITION 16.7. *Suppose* $\chi_1, \ldots, \chi_k$ *are all the constraints on* $\mathbb{K}$*, and these constrain sets* $A_1, \ldots, A_k$ *respectively. Then*

1. $A_1, \ldots, A_k$ *are relatively free in* $\mathbb{K}$*, and*
2. $B = \mathcal{L} \backslash \bigcup_{i=1}^{k} A_i$ *is jointly free in* $\mathbb{K}$.

Proof. Part 1. We are given $p \in \mathbb{K}$ and $x \in [0,1]^{2^{|A|}}$ such that for all $i = 1, \ldots, k$ and $\alpha_i \in \mathcal{A}A_i$, $\sum_{\alpha \sim \alpha_i} x_\alpha = p(\alpha_i)$.

Now $p \in \mathbb{K}$ if and only if for each constraint (i.e., for $i = 1, \ldots, k$), $f_i(p(\alpha_{i,1}), \ldots, p(\alpha_{i,2^{l_i}})) = 0$.

Take any probability function q on $\mathcal{L}$ such that for all $\alpha \in \mathcal{A}A$, $q(\alpha) = x_\alpha$. We are ensured that $q(\alpha_{i,j}) = p(\alpha_{i,j})$ for each constraint ($i = 1, \ldots, k$) and atom of the set that it constrains ($j = 1, \ldots, 2^{l_i}$).

Consequently $f_i(q(\alpha_{i,1}), \ldots, q(\alpha_{i,2^{l_i}})) = 0$ for $i = 1, \ldots, k$, and $q \in \mathbb{K}$.

---

and any $x \in [0,1]$ there is some $q \in \mathbb{K}$ such that $q(\theta \mid \phi) = x$ yet $q = p$ everywhere else. In the light of this condition, we do not need the qualification that $p(\beta) > 0$ in part 5 of this proposition, for $\exists p \in \mathbb{K}, \forall \alpha \forall \beta, p(\beta) > 0 \Rightarrow p(\alpha \mid \beta) = y_{\alpha\beta}$ if and only if $\exists q \in \mathbb{K}, \forall \alpha \forall \beta, q(\alpha \mid \beta) = y_{\alpha\beta}$.

[⋆] This disjointness condition is not, in fact, necessary. However, we do not need the more general result here, and the disjointness condition renders the proof straightforward.

Part 2. Since $\chi_1, \ldots, \chi_k$ are all the constraints on $\mathbb{K}$, none of the variables in B are constrained at all. Any probability function q which agrees with given $p \in \mathbb{K}$ on $A = \bigcup_{i=1}^{k} A_i$ will satisfy all the constraints and so will be in $\mathbb{K}$. □

It is not hard to see that, if there are no constraints in operation, cross entropy distance from p is minimised by p itself, and if p maximises entropy then it renders all variables probabilistically independent. We can generalise this: if constraints are in operation, cross entropy distance is minimised by a function that agrees with p as much as possible, and if p maximises entropy then it satisfies as many conditional independencies as the constraints allow. The following lemma makes these notions precise.

LEMMA 16.8.

1. *Suppose we are given two probability functions p, q defined over the same finite domain $\mathcal{L} = A \cup B$, where A and B are disjoint. Define the probability function r over $\mathcal{L}$ by $r(\alpha\beta) = p(\beta \mid \alpha)q(\alpha)$ (for all $\alpha \in \mathcal{A}A$, $\beta \in \mathcal{A}B$). Then $d_{\mathcal{L}}(r, p) \leq d_{\mathcal{L}}(q, p)$ (with equality if $r = q$).*
2. *Suppose p is defined over $\mathcal{L} = A \cup B \cup C$, where A, B and C are disjoint. Define q over $\mathcal{L}$ by $q(\alpha\beta\gamma) = p(\gamma \mid \alpha)p(\alpha\beta)$ (for all $\alpha \in \mathcal{A}A$, $\beta \in \mathcal{A}B$, $\gamma \in \mathcal{A}C$). Then $H_{\mathcal{L}}(q) \geq H_{\mathcal{L}}(p)$ with equality iff $I_p(B, C \mid A)$.*

Proof. Part 1.

$$\begin{aligned} d_{\mathcal{L}}(q, p) &= \sum_{\alpha,\beta} q(\alpha\beta) \log \frac{q(\alpha\beta)}{p(\alpha\beta)} \\ &= \sum_{\alpha,\beta} q(\beta \mid \alpha)q(\alpha) \log \frac{q(\beta \mid \alpha)q(\alpha)}{p(\beta \mid \alpha)p(\alpha)} \\ &= \sum_{\alpha,\beta} q(\beta \mid \alpha)q(\alpha) \left[\log \frac{q(\beta \mid \alpha)}{p(\beta \mid \alpha)} + \log \frac{q(\alpha)}{p(\alpha)}\right] \\ &= \sum_{\alpha,\beta} q(\beta \mid \alpha)q(\alpha) \log \frac{q(\beta \mid \alpha)}{p(\beta \mid \alpha)} + \sum_{\alpha} q(\alpha) \log \frac{q(\alpha)}{p(\alpha)}. \end{aligned}$$

Likewise,

$$\begin{aligned} d_{\mathcal{L}}(r, p) &= \sum_{\alpha,\beta} p(\beta \mid \alpha)q(\alpha) \log \frac{p(\beta \mid \alpha)}{p(\beta \mid \alpha)} + \sum_{\alpha} q(\alpha) \log \frac{q(\alpha)}{p(\alpha)} \\ &= \sum_{\alpha} q(\alpha) \log \frac{q(\alpha)}{p(\alpha)}. \end{aligned}$$

This is no greater than $d_{\mathcal{L}}(q, p)$, and is smaller unless $q(\beta \mid \alpha) = p(\beta \mid \alpha)$ for each α, β, or $p(\alpha) = 0 \neq q(\alpha)$ for some α.*

Part 2.

$$\begin{aligned}
H_{\mathcal{L}}(q) - H_{\mathcal{L}}(p) &= -\sum_{\alpha\beta} p(\alpha\beta) \log p(\alpha\beta) - \sum_{\alpha,\beta,\gamma} p(\gamma \mid \alpha) p(\alpha\beta) \log p(\gamma \mid \alpha) \\
&\quad + \sum_{\alpha\beta} p(\alpha\beta) \log p(\alpha\beta) \\
&\quad + \sum_{\alpha,\beta,\gamma} p(\gamma \mid \alpha\beta) p(\alpha\beta) \log p(\gamma \mid \alpha\beta) \\
&= \sum_{\alpha,\beta,\gamma} p(\gamma \mid \alpha\beta) p(\alpha\beta) \log p(\gamma \mid \alpha\beta) \\
&\quad - \sum_{\alpha,\gamma} p(\gamma \mid \alpha) p(\alpha) \log p(\gamma \mid \alpha) \\
&= \sum_{\alpha,\beta,\gamma} p(\gamma \mid \alpha\beta) p(\alpha\beta) \log p(\gamma \mid \alpha\beta) \\
&\quad - \sum_{\alpha,\gamma} \left[\sum_{\beta} p(\gamma \mid \alpha\beta) p(\beta \mid \alpha) \right] p(\alpha) \log p(\gamma \mid \alpha) \\
&= \sum_{\alpha,\beta,\gamma} p(\gamma \mid \alpha\beta) p(\alpha\beta) \log \frac{p(\gamma \mid \alpha\beta)}{p(\gamma \mid \alpha)} \\
&= \sum_{\alpha,\beta,\gamma} p(\alpha\beta\gamma) \log \frac{p(\alpha\beta\gamma)}{p(\gamma \mid \alpha) p(\alpha\beta)} \\
&= d_{\mathcal{L}}(p, q) \geq 0
\end{aligned}$$

with equality $\Leftrightarrow p = q \Leftrightarrow p(\gamma \mid \alpha\beta) = p(\gamma \mid \alpha)$ for all $\alpha, \beta, \gamma \Leftrightarrow I_p(B, C \mid A)$, as required. □

THEOREM 16.9. *Suppose*

- $\chi_1, \ldots, \chi_k$ *are all the constraints on* $\mathbb{K}$, *and these constrain the sets* $A_1, \ldots, A_k$ *respectively. Let* $A = \bigcup_{i=1}^{k} A_i$ *and* $B = \mathcal{L} \backslash A$, *and for any set* $X \subseteq \mathcal{L}$ *let* $X_0 = X \cap \mathcal{L}_0$ *and* $X_+ = X \cap \mathcal{L}_+$.

* This second possibility can be ignored if the open-mindedness condition is adopted – see Section 10.

– *The constraint sets* $A_1, \ldots, A_k$ *do not intersect in* $\mathcal{L}_+$,* *and*
– p *is determined by the entropic assignment* ρ_e.
► *Then*

1. $p(\beta \mid \alpha) = p_0(\beta \mid \alpha)$ *for each* $\alpha \in \mathcal{A}A_0, \beta \in \mathcal{A}B_0$.
2. $I_p(A_{i+}, \mathcal{L}\backslash A_i \mid A_{i0})$ *for* $i = 1, \ldots, k$.

Proof. Part 1. $A_1, \ldots, A_k$ are relatively free in $\mathbb{K}$ with respect to $\mathcal{L}$ (Proposition 16.7.1). Therefore $A_{10}, \ldots, A_{k0}$ are relatively free in $\mathbb{K}_0$ with respect to $\mathcal{L}_0$ (Proposition 16.6.2), and B is jointly free in $\mathbb{K}_0$ with respect to $\mathcal{L}_0$ (Proposition 16.7.2).

Now by Proposition 16.4.5 and its footnote, there is a $q \in \mathbb{K}_0$ such that $q(\beta \mid \alpha) = p_0(\beta \mid \alpha)$ for all $\alpha \in \mathcal{A}A_0, \beta \in \mathcal{A}B_0$, and so by Lemma 16.8.1 any function in $\mathbb{K}_0$ that minimises cross entropy must have this property. Since functions in $\mathbb{K}_0$ are just restrictions of functions in $\mathbb{K}$, any function in $\mathbb{K}$ that minimises cross entropy must have this property, and so must p.

Part 2. The strategy is similar here. We need to show that there is a $q \in \mathbb{K}$ such that $I_q(A_{i+}, \mathcal{L}\backslash A_i \mid A_{i0})$ for $i = 1, \ldots, k$. Then it follows by Lemma 16.8.2 that any function that maximises entropy must have this property, and so p must have the property.

Take an arbitrary $r \in \mathbb{K}$ and define $x \in [0,1]^{2^{|A|}}$ by setting $x_\alpha = r(A_0^\alpha) \prod_{i=1}^k r(A_{i+}^\alpha \mid A_{i0}^\alpha)$ for each $\alpha \in \mathcal{A}A$. (If $A_{i+}^\alpha = \emptyset$ then we take $r(A_{i+}^\alpha \mid A_{i0}^\alpha) = 1$.) Now for $\alpha_j \in \mathcal{A}A_j$,

$$\begin{aligned}
\sum_{\alpha \sim \alpha_j} x_\alpha &= \sum_{\alpha \sim \alpha_j} r(A_0^\alpha) \prod_{i=1}^k r(A_{i+}^\alpha \mid A_{i0}^\alpha) \\
&= \sum_{\alpha \sim \alpha_j} r(A_0^\alpha) r(A_{j+}^\alpha \mid A_{j0}^\alpha) \\
&= r(A_{j_0}^{\alpha_j}) r(A_{j+}^{\alpha_j} \mid A_{j0}^{\alpha_j}) \\
&= r(A_j^{\alpha_j}) \\
&= r(\alpha_j).
\end{aligned}$$

(Note that the second equality above requires the condition that the A_{i+} be disjoint.)

Since $A_1, \ldots, A_k$ are relatively free (Proposition 16.7.1) there is a $q \in \mathbb{K}$ such that $q(\alpha) = x_\alpha$ for all $\alpha \in \mathcal{A}A$. By construction we have that $I_q(A_{i+}, \mathcal{L}\backslash A_i \mid A_{i0})$, as required. □

* This corresponds to the condition of Theorem 14.1 that all partners of the new variables are in $\mathcal{L}_0$. Without loss of generality we assume here that each $c \in \mathcal{L}_+$ occurs in no more than one constraint, since we can combine constraints in which c occurs to ensure that this is so.

Theorem 14.1 is a straightforward consequence of the above result. Thus Theorem 16.9 generalises Theorem 14.1 to handle non-linear constraints.

17. Conclusion

We have seen how Bayesianism can begin to tackle the problem of language change. The first step is to escape from the blinkers of language invariance arguments: degrees of belief must change as language changes, but they need only change conservatively. Then one can give a practical procedure for changing degrees of belief. The procedure developed here – the entropic assignment – generalises Bayesian conditionalisation to determine an agent's new probability function given her old function, her new knowledge, *and her new language*. Not all Bayesians will be comfortable with routine use of strong constraints like the maximum entropy principle, but those who are will no doubt welcome the possibility of efficient updating by using Bayesian networks.

By addressing the problem of language change we open a whole can of worms. How can transitional knowledge be made explicit and quantified? How does the entropic assignment compare with other possible rational assignments that one might put forward? Can one further relax the conditions under which the Bayesian network representation is applicable? How do we generalise to more realistic languages? If we dissect these worms then the pieces will no doubt grow into new worms, but Bayesianism will be rendered more applicable and more widely testable.

Acknowledgements

Thanks to an anonymous referee, David Corfield and Donald Gillies for useful comments, and to the U.K. Arts and Humanities Research Board for supporting this research.

References

Binder, J., Koller, D., Russell, S., and Kanazawa, K., 1997, "Adaptive probabilistic networks with hidden variables," *Machine Learning* **29**, 213–244.

Carnap, R., 1950, *Logical Foundations of Probability*, London: Routledge & Kegan Paul (second edition 1962).

Carnap, R., 1971, "A basic system of inductive logic part 1," pp. 33–165 in *Studies in Inductive Logic and Probability*, Vol. 1, R. Carnap and R.C. Jeffrey, eds., Berkeley, CA: University of California Press.

Christensen, D., 2000, "Diachronic coherence versus epistemic impartiality," *The Philosophical Review* **109**, 349–371.

Corfield, D., 2001, "Bayesianism in mathematics," pp. 175–201 in *Foundations of Bayesianism*, D. Corfield and J. Williamson, eds., Dordrecht: Kluwer Academic Publishers.

Cowell, R.G., Dawid, A.P., Lauritzen, S.L., and Spiegelhalter, D.J., 1999, *Probabilistic Networks and Expert Systems*, Berlin: Springer-Verlag.

Cristianini, N. and Shawe-Taylor, J., 2000, *Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge: Cambridge University Press.

Earman, J., 1992, *Bayes or Bust?*, Cambridge, MA: MIT Press.

de Finetti, B., 1937, "Foresight. Its logical laws, its subjective sources," pp. 53–118 in *Studies in Subjective Probability*, H.E. Kyburg and H.E. Smokler, eds., London: Wiley [1964].

Forster, M.R., 1995, "Bayes and bust: Simplicity as a problem for a probabilist's approach to confirmation," *British Journal for the Philosophy of Science* **46**, 399–324.

Forster, M. and Sober, E., 1994, "How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions," *British Journal for the Philosophy of Science* **45**, 1–35.

Frege, G., 1880, "Boole's logical calculus and the concept-script," pp. 9–52 in *Posthumous Writings – Gottlob Frege*, H. Hermes, F. Kambartel, and F. Kaulbach, eds., Oxford: Blackwell [1979].

Gaifman, H. and Snir, M., 1982, "Probabilities over rich languages," *Journal of Symbolic Logic* **47**, 495–548.

Gärdenfors, P., 1990, "The dynamics of belief systems: foundations versus coherence theories," *Revue Internationale de Philosophie* **44**, 24–46.

Gillies, D., 1991, "Intersubjective probability and confirmation theory," *British Journal for the Philosophy of Science* **42**, 513–533.

Gillies, D., 1996, *Artificial Intelligence and Scientific Method*, Oxford: Oxford University Press.

Gillies, D., 2001, "Bayesianism and the fixity of the theoretical framework," pp. 363–379 in *Foundations of Bayesianism*, D. Corfield and J. Williamson, eds., Dordrecht: Kluwer Academic Publishers.

Goldstick, D., 1971, "Methodological conservatism," *American Philosophical Quarterly* **8**, 186–191.

Goodman, N., 1954, *Fact, Fiction and Forecast*, fourth edition, Harvard: Harvard University Press.

Halpern, J.Y. and Koller, D., 1995, "Representation dependence in probabilistic inference," pp. 1853–1860 in *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI 95)*, C.S. Mellish, ed., San Francisco, CA: Morgan Kaufmann Publishers.

Harman, G., 1986, *Change in View: Principles of Reasoning*, Cambridge, MA: MIT Press.

Hausser, R., 1999, *Foundations of Computational Linguistics*, Berlin: Springer-Verlag.

Hilpinen, R., 1975, "Carnap"s new system of inductive logic," pp. 333–359 in *Rudolf Carnap, Logical Empiricist: Materials and Perspectives*, J. Hintikka, ed., Dordrecht: D. Reidel.

Howson, C., 1976, "The development of logical probability," pp. 277–298 in *Essays in Memory of Imre Lakatos*, R.S. Cohen, P.K. Feyerabend, and M.W. Wartofsky, eds., Boston Studies in the Philosophy of Science, Dordrecht: D. Reidel.

Howson, C., 1997, "Bayesian rules of updating," *Erkenntnis* **45**, 195–208.

Howson, C., 2001, "The logical basis of uncertainty," pp. 137–159 in *Foundations of Bayesianism*, D. Corfield and J. Williamson, eds., Dordrecht: Kluwer Academic Publishers.

Howson, C. and Peter Urbach, P., 1989, *Scientific Reasoning: The Bayesian Approach*, Chicago, IL: Open Court (second edition, 1993).

James, W., 1907, "What pragmatism means," pp. 141–158 in *Essays in Pragmatism by William James*, A. Castell, ed., New York: Hafner [1948].

Jaynes, E.T., 1973, "The well-posed problem," *Foundations of Physics* **3**, 477–492.

Jaynes, E.T., 1998, "Probability theory: The logic of science," http:// bayes.wustl.edu/

Jim, K.-C. and Giles, C.L., 2000, "Talking helps: Evolving communicating agents for the predator-prey pursuit problem," *Artificial Life* **6**, 237–254.

Kakas, A.C., Kowalski, R., and Toni, F., 1998, "The role of abduction in logic programming," pp. 235–324 in *Handbook of Logic in Artificial Intelligence and Logic Programming 5*, D.M. Gabbay, C.J. Hogger, and J.A. Robinson, eds., Oxford: Oxford University Press.

Kuhn, T.S., 1962, *The Structure of Scientific Revolutions*, Chicago, IL: University of Chicago Press (second edition, 1970).

Kvasz, L., 2000, "Changes of language in the development of mathematics," *Philosophia Mathematica* **8**, 47–83.

Kwoh, C.-K. and Gillies, D.F., 1996, "Using hidden nodes in Bayesian networks," *Artificial Intelligence* **88**, 1–38.

Lakatos, I., 1968, "Changes in the problem of inductive logic," pp. 315–417 in *The Problem of Inductive Logic, Proceedings of the International Colloquium in the Philosophy of Science*, London, 1965, Vol. 2, I. Lakatos, ed., Amsterdam: North-Holland.

Laudan, L., 1981, "A confutation of convergent realism," *Philosophy of Science* **48**, 19–48.

Lauritzen, S.L. and Spiegelhalter, D.J., 1988, "Local computation with probabilities in graphical structures and their applications to expert systems," (with discussion) *Journal of the Royal Statistical Society B* **50**, 157–254.

Lehrer, K., 1974, *Knowledge*, Oxford: Clarendon Press.

Lehrer, K., 1978, "Why not scepticism?," pp. 346–363 in *Essays on Knowledge and Justification*, G. Pappas and M. Swain, eds., Ithaca, NY: Cornell University Press.

Lycan, W.G., 1988, *Judgement and Justification*, Cambridge: Cambridge University Press.

Magnani, L., 2001, *Abduction, Reason, and Science: Processes of Discovery and Explanation*, Dordrecht: Kluwer Academic Publishers/Plenum Publishers.

Muggleton, S. and de Raedt, L., 1994, "Inductive logic programming: theory and methods," *Journal of Logic Programming* **19–20**, 629–679.

Nagel, E., 1963, "Carnap's theory of induction," pp. 785–825 in *The Philosophy of Rudolf Carnap*, P.A. Schilpp, ed., Chicago, IL: Open Court.

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

Paris, J.B., 1994, *The Uncertain Reasoner's Companion*, Cambridge: Cambridge University Press.

Paris, J.B. and Vencovská, A., 1997, "In defense of the maximum entropy inference process," *International Journal of Automated Reasoning* **17**, 77–103.

Paris, J.B. and Vencovská, A., 2001, "Common sense and stochastic independence," pp. 203–240 in *Foundations of Bayesianism*, D. Corfield and J. Williamson, eds., Dordrecht: Kluwer Academic Publishers.

Pearl, J., 1988, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Mateo, CA: Morgan Kaufmann.

Polya, G., 1954, *Patterns of Plausible Inference*, Mathematics and Plausible Reasoning, Vol. 2, Princeton, NJ: Princeton University Press.

Putnam, H., 1963, "'Degree of confirmation' and inductive logic," pp. 761–783 in *The Philosophy of Rudolf Carnap*, P.A. Schilpp, ed., Chicago, IL: Open Court.

Quine, W.V.O., 1960, *Word and Object*, Cambridge, MA: MIT Press and New York: John Wiley.

Ramsey, F.P., 1926, "Truth and probability," pp. 23–52 in *Studies in Subjective Probability*, H.E. Kyburg and H.E. Smokler, eds., London: Wiley [1964].

Rosenkrantz, R.D., 1977, *Inference, Method and Decision: Towards a Bayesian Philosophy of Science*, Dordrecht: D. Reidel.

Ross, L. and Anderson, C.A., 1982, "Shortcoming in the attribution process: on the origins and maintenance of erroneous social assessments," pp. 129–152 in *Judgements under Uncertainty: Heuristics and Biases*, D. Kahneman, P. Slovic, and A. Tversky, eds., Cambridge: Cambridge University Press.

Rott, H., 1999, "Coherence and conservatism in the dynamics of belief," *Erkenntnis* **50**, 387–412.

Shimony, A., 1955, "Coherence and the axioms of confirmation," *Journal of Symbolic Logic* **20**, 1–28.

Sklar, L., 1975, "Methodological conservatism," *Philosophical Review* **84**, 374–400.

Sober, E., 1975, *Simplicity*, Oxford: Oxford University Press.

Spirtes, P., Glymour, C., and Scheines, R., 1993, *Causation, Prediction, and Search*, Cambridge, MA: MIT Press (second edition, 2000).

Thagard, P., 1988, *Computational Philosophy of Science*, Cambridge, MA: MIT Press.

Vapnik, V.N., 1995, *The Nature of Statistical Learning Theory*, Berlin: Springer-Verlag (second edition, 2000).

Williams, P.M., 1980, "Bayesian conditionalisation and the principle of minimum information," *British Journal for the Philosophy of Science* **31**, 131–144.

Williamson, J., 1999, "Countable additivity and subjective probability," *British Journal for the Philosophy of Science* **50**, 401–416.

Williamson, J., 2001a, "Abduction and its distinctions," Review of "Abduction, Reason, and Science: Processes of Discovery and Explanation" by L. Magnani, *British Journal for the Philosophy of Science*, to appear.

Williamson, J., 2001b, "Machine Learning and the Philosophy of Science: a Dynamic Interaction," pp. 1–12 in *Proceedings of the ECML-PKDD-01 Workshop on Machine Learning as Experimental Philosophy of Science*, Freiburg, K. Korb and H. Bensusan, eds.

Williamson, J., 2002a, "Probability logic," pp. 397–424 in *Handbook of the Logic of Argument and Inference: The Turn Toward the Practical*, D. Gabbay, R. Johnson, H.J. Ohlbach, and J. Woods, eds., Amsterdam: Elsevier.

Williamson, J., 2002b, "Maximising entropy efficiency," *Electronic Transactions in Artificial Intelligence*, to appear.