Linköping Electronic Articles in
Computer and Information Science
Vol. 7(2002): nr 0

Maximising Entropy Efficiently

Jon Williamson

Department of Philosophy
King's College, London, UK

Linköping University Electronic Press
Linköping, Sweden

http://www.ep.liu.se/ea/cis/2002/00/

Published on September 18, 2002 by
Linköping University Electronic Press
581 83 Linköping, Sweden

Linköping Electronic Articles in
Computer and Information Science
ISSN 1401-9841
Series editor: Erik Sandewall

Typeset by the author using LaTeX
Formatted using étendu *style*

Recommended citation:
<Author>. <Title>. Linköping Electronic Articles in Computer and Information Science, Vol. 7(2002): nr 0. http://www.ep.liu.se/ea/cis/2002/00/. *September 18, 2002.*

This URL will also contain a link to the author's home page.

Abstract

Determining a prior probability function via the maximum entropy principle can be a computationally intractable task. However one can easily determine — in advance of entropy maximisation — a list of probabilistic independencies that the maximum entropy function will satisfy. These independencies can be used to reduce the complexity of the entropy maximisation task. In particular, one can use these independencies to construct a directed acyclic graph in a Bayesian network, and then maximise entropy with respect to the numerical parameters of this network. This can result in an efficient representation of a prior probability function, and one that may allow efficient updating and marginalisation. The computational complexity of maximising entropy can be further reduced when knowledge of causal relationships is available. Moreover, the proposed simplification of the entropy maximisation task may be exploited to construct a proof theory for probabilistic logic.

1 Introduction

Bayesians argue that an agent's degrees of belief ought to respect the axioms of probability: a belief function ought to be a probability function. Many Bayesians impose further mechanisms for choosing a belief function, given an agent's background knowledge. In particular, many accept the *maximum entropy principle*, which says that one ought to adopt, out of all the probability functions that satisfy the constraints imposed by background knowledge, a function p that maximises entropy

$$H = -\sum_{v} p(v) \log p(v) \qquad (1)$$

(where the sum is taken over the assignments v to elements of the domain — the notation will be explained in §2). This principle is often justified on the grounds that a maximum entropy probability function is a *least prejudiced* probability function that satisfies the constraints. It satisfies the constraints but commits to as little as possible beyond the knowledge embodied by the constraints.[1] A second justification cites a number of intuitively plausible conditions that any principle for determining a probability function from background knowledge ought to satisfy, and goes on to show that the maximum entropy principle is the only principle which satisfies these conditions.[2] The maximum entropy principle is still a matter of some controversy,[3] but I shall not dwell on the issue of justification any further here.

The chief difficulty for those who do accept the maximum entropy principle is that the number of parameters $p(v)$ in the entropy expression (Equation 1) is exponential in the size of the domain (see §3), so when the domain size is large it can be impractical to determine the values of the parameters that maximise entropy. The object of this paper is to put forward a principled and practical way of reducing the number of parameters required in the entropy maximisation process.

The key idea is this. By analysing the structure of the constraints imposed by background knowledge, it is possible to determine a host of conditional probabilistic independencies that the maximum entropy probability function p will satisfy. In §4 we shall see that the independence structure of p is most naturally represented by a Markov network. By transforming this Markov network into a Bayesian network (§5), we can exploit these independencies to reparameterise the entropy expression, thereby reducing the computational complexity of the maximisation task.

Apart from simplifying the entropy maximisation problem, exploiting the independencies inherent in the maximum entropy function yields the following advantages. First, we are left with a Bayesian network representation of an agent's belief function: this is desirable

[1] [Jaynes 1957].
[2] [Paris & Vencovská 2001].
[3] See [Halpern & Koller 1995], [Paris & Vencovská 1997].

in that it may allow efficient storage and updating of the belief function (§5). Second, the approach allows further computational savings when the background knowledge includes knowledge of conditional independencies or causal relationships (§6). Third, the approach can be extended to cope with reasoning in domains which have logical as well as probabilistic structure (§7).

2 Framework and Notation

We shall be concerned with a finite domain $V = \{A_1, \ldots, A_n\}$ of discrete variables. Each variable A_i can take one of K_i possible values, and the expression $a_i@A_i$ signifies that a_i is the assignment of A_i to one of its values. For a subset $U = \{A_{i_1}, \ldots, A_{i_k}\}$ of V, an assignment $u@U$ consists of an assignment to each of the variables in U, and u may be written $a_{i_1} \ldots a_{i_k}$, where $a_{i_1}@A_{i_1}, \ldots, a_{i_k}@A_{i_k}$. Thus upper-case letters refer to variables or sets of variables, while the corresponding lower-case letters refer to their assignments. Two assignments $u_1@U_1$ and $u_2@U_2$ are *consistent*, written $u_1 \sim u_2$, if they agree on $U_1 \cap U_2$.

A probability function p over V is determined by its values on the parameters $x^v =_{df} p(v)$, where v ranges over the assignments to V.[4] Each $x^v \in [0,1]$ and by additivity of probability $\sum_{v@V} x^v = 1$. Given some fixed ordering of assignments to V, let x denote the vector of parameters $(x^v)_{v@V}$. The set of probability functions then corresponds to the space $\mathbb{P} = \{x \in [0,1]^J : \sum_{v@V} x^v = 1\}$, where $J = \prod_{i=1}^n K_i$, the number of assignments to V.

An agent's background knowledge is assumed to impose a number of constraints $\chi_1, \ldots, \chi_m$ on the set of probability functions that she may adopt. Associated with each constraint χ_i is the set C_i of variables involved in the constraint: if, for example, χ_i is the constraint that the mean of variable A_1 is $1/3$ then the associated constraint set is $C_i = \{A_1\}$. Let $z_i^{c_i} =_{df} p(c_i)$ where $c_i@C_i$, and let z_i be the vector of these parameters. Each constraint χ_i on C_i will be assumed to be an equality constraint of the form $f_i(z_i) = 0$ or an inequality constraint of the form $f_i(z_i) \geq 0$. Note that z_i is determined by x through the relationship $z_i^{c_i} = \sum_{v@V, v\sim c_i} x^v$. Denote the set of constrained probability functions by $\mathbb{C}$, so

$$\mathbb{C} = \{x \in \mathbb{P} : f_1(z_1) \geqq 0, \ldots, f_m(z_m) \geqq 0\},$$

where $\geqq$ is either $\geq$ or $=$ according to the constraint. We shall assume throughout that the constraints $\chi_1, \ldots, \chi_m$ are consistent in the sense that $\mathbb{C} \neq \emptyset$, since maximising entropy subject to inconsistent constraints is trivial.

[4] [Paris 1994] pp. 13-14.

3 Maximising Entropy

Under the above x-parameterisation, the entropy equation is

$$H = -\sum_{v@V} x^v \log x^v \tag{2}$$

The maximum entropy principle requires that a parameter vector $x \in \mathbb{C}$ is found that maximises H. Typically one might use numerical methods such as gradient ascent to adjust the x^v until a local maximum is found or one might determine a local maximum by using the method of Lagrange multipliers. One needs of course to be wary of the following possibilities: (i) there may be no global maximum (although if the constraints limit x to a closed subset $\mathbb{C}$ of $\mathbb{P}$, then a global maximum will exist); (ii) there may be more than one global maximum (in which case Bayesians may deem the agent in question to be rational whichever global maximum she chooses as her belief function); (iii) the above methods may find a local maximum that is not the global maximum (note that there are no non-global maxima if $\mathbb{C}$ is convex, which occurs for example if the constraint functions f_i are all linear[5]).

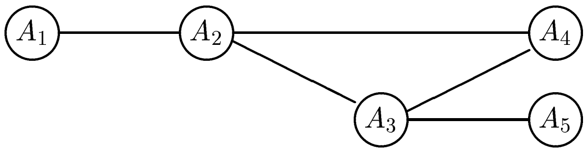
Perhaps the most serious difficulty is this. There are $\prod_{i=1}^n K_i$ assignments on V. One of the x-parameters is determined by additivity from the others, and so there are $(\prod_{i=1}^n K_i) - 1$ free x-parameters, a number exponential in n. This is a problem for numerical methods because as n becomes large there will quickly become too many parameters to be stored and adjusted, and there may even be too many terms in Equation 2 to be summed in available time. Lagrange multiplier methods suffer analogously: a system of equations (consisting of the m constraint equations and $\prod_{i=1}^n K_i$ partial derivatives of the Lagrange equation with respect to the x-parameters) must be solved for x, and this system of equations will quickly become unhandleable as n increases.

Unfortunately there appears to be no fully general solution to the complexity problem: the task of finding an approximation to the maximum entropy function is NP-complete[6] and the task of finding a likely approximation is RP-complete,[7] and so if NP $\neq$P $\neq$ RP then there is no polynomial time algorithm for performing these tasks and any algorithm will be intractable in the worst case as n increases. The best we can hope for is an algorithm which performs well on the type of problem that occurs in practice and badly only rarely. This at least would be an improvement on naive numerical and Lagrange multiplier approaches which perform uniformly badly.

The approach outlined in the rest of this paper is based on the premise that in practice the sizes of the constraint sets C_i are usually small in comparison with n, as n becomes large. Constraints often

[5][Paris 1994] Proposition 6.1.
[6][Paris 1994] Theorem 10.6.
[7][Paris 1994] Theorem 10.7.

Figure 1: Example constraint graph



consist of observed means of single variables, marginals of small sets of variables, hypothesised deterministic connections amongst small sets of variables, causal connections amongst pairs of variables, independence relationships amongst small sets of variables, and so on. The point is that there is a limit to the amount we normally observe and to the connections amongst variables posited by background knowledge, in that while there may be many observations and many connections, each observation and connection will relate only few variables. The number of possible observations pertinent to a joint distribution over V increases exponentially with n, but, I suggest, our ability to observe increases sub-exponentially.

If such an assumption is correct, then as n grows there are many conditional independencies that the entropy-maximising probability function p will satisfy. We can identify these independencies just from the constraint sets C_i, and exploit them to simplify the task of determining p, as we shall now see.

4 From Constraints to Markov Network

Define an undirected *constraint graph* $\mathcal{G}$ as follows. Take as vertices the variables in V. Include an edge between two variables $A_i, A_j \in V$ if and only if A_i and A_j occur in the same constraint set C_k.

Suppose, for example, that $V = \{A_1, \ldots, A_5\}$ and that there are four constraints $\chi_1, \ldots, \chi_4$ constraining $C_1 = \{A_1, A_2\}, C_2 = \{A_2, A_3, A_4\}, C_3 = \{A_3, A_5\}, C_4 = \{A_4\}$ respectively. Then the constraint graph $\mathcal{G}$ is depicted in Figure 1.

The constraint graph is useful because it represents conditional independencies that a maximum entropy function p satisfies. This is best explained with the help of some additional terminology. Given disjoint $X, Y, Z \subseteq V$, we shall write $X \perp\!\!\!\perp_p Y \mid Z$ if, according to a probability function p, X is probabilistically independent of Y conditional on Z, and we shall write $X \not\perp\!\!\!\perp_p Y \mid Z$ if this independence does not hold, i.e. if X and Y are probabilistically dependent conditional on Z. We say that Z *separates* X from Y in undirected graph $\mathcal{G}$ if every path from a vertex in X to a vertex in Y goes through some vertex in Z.

Theorem 4.1 If Z separates X from Y in the constraint graph $\mathcal{G}$ then $X \perp\!\!\!\perp_p Y \mid Z$ for any p satisfying the constraints which maximises entropy.

Proof: The first step is to use standard Lagrange multiplier optimisation. By theorems of Lagrange and Runge-Kutta,[8] if $x \in \mathbb{C}$ is a local maximum of H then there are constants $\mu, \lambda_1, \ldots, \lambda_m \in \mathbb{R}$, called *multipliers*, such that

$$\frac{\partial H}{\partial x^v} + \mu + \sum_{i=1}^{m} \lambda_i \frac{\partial f_i}{\partial x^v} = 0 \tag{3}$$

for each assignment $v@V$, where μ is the multiplier corresponding to the additivity constraint $\sum_{v@V} x^v = 1$, and where $\lambda_i = 0$ for each inequality constraint which is not effective at x (i.e. for each inequality constraint χ_i such that $f_i(x) > 0$).

Now the argument of f_i is the vector z_i of probabilities of assignments to C_i. Moreover, $z_i^{c_i} = \sum_{v@V, v \sim c_i} x^v$, so

$$\frac{\partial f_i}{\partial x^v} = \frac{\partial f_i}{\partial z_i^{c_i}} \frac{\partial z_i^{c_i}}{\partial x^v} = \frac{\partial f_i}{\partial z_i^{c_i}} .1$$

where c_i is the assignment to C_i that is consistent with v. Furthermore,

$$\frac{\partial H}{\partial x^v} = -1 - \log x^v,$$

so Equation 3 can be written

$$\log x^v = -1 + \mu + \sum_{i=1}^{m} \lambda_i \frac{\partial f_i}{\partial z_i^{c_i}}$$

where each $c_i \sim v$. Thus,

$$x^v = e^{\mu - 1} \prod_{i=1}^{m} e^{\lambda_i \frac{\partial f_i}{\partial z_i^{c_i}}} \tag{4}$$

Hence the local maximum x is representable as a product of functions, each of which depends only on variables in a single constraint set C_i (the leading term is a constant). The probability function p corresponding to x is said to *factorise* according to the constraint sets $C_1, \ldots C_m$, and since these sets are complete subsets of $\mathcal{G}$, p is said to *factorise* according to $\mathcal{G}$.[9] The *global Markov condition* says that if Z separates X from Y in $\mathcal{G}$ then $X \perp\!\!\!\perp_p Y \mid Z$, and this condition is a straightforward consequence of factorisation according to $\mathcal{G}$.[10] Thus the theorem follows for local maxima p, and in particular for global maxima p. ■

The converse does not hold in general. For example, a constraint χ_1 that asserts the independence of A_1 and A_2 must of course be

[8] See for example [Sundaram 1996], Theorems 5.1 and 6.1.
[9] [Lauritzen 1996] 34-35.
[10] [Lauritzen 1996] Proposition 3.8.

satisfied by the maximum entropy function p, but would not correspond to any separation in the constraint graph $\mathcal{G}$. However, there is a partial converse to Theorem 4.1: separation in $\mathcal{G}$ captures all the conditional independencies of p that are due to structure of the constraint sets and not the constraints themselves. More precisely, suppose that as before we are given disjoint $X, Y, Z \subseteq V$ and constraint sets $C_1, \ldots, C_m$ and we construct the corresponding constraint graph $\mathcal{G}$; then

Theorem 4.2 If, for all $\chi_1, \ldots, \chi_m$ constraining $C_1, \ldots, C_m$ respectively, $X \perp\!\!\!\perp_p Y \mid Z$ where p is a function satisfying $\chi_1, \ldots, \chi_m$ that maximises entropy, then Z separates X from Y in $\mathcal{G}$.

Proof: We shall show the contrapositive, namely that if Z does not separate X from Y in $\mathcal{G}$ then there is some $\chi_1, \ldots, \chi_m$ constraining $C_1, \ldots, C_m$ such that, for p a maximum entropy satisfier of $\chi_1, \ldots, \chi_m$, $X \rightleftharpoons_p Y \mid Z$.

So suppose $A_{i_1}, \ldots, A_{i_k}$ is a shortest path from some $A_{i_1} \in X$ to some $A_{i_k} \in Y$ avoiding vertices in Z. The task is then to find some $\chi_1, \ldots, \chi_m$ that render A_{i_1} and A_{i_k} probabilistically dependent conditional on Z for the maximum entropy p.

For $j = 1, \ldots, k-1$, A_{i_j} and $A_{i_{j+1}}$ are connected by an edge in $\mathcal{G}$, so they are in the same constraint set, which we can call C_j without loss of generality. Moreover no three vertices on the path are in the same constraint set, for we could otherwise construct a shorter path from A_{i_1} to A_{i_k} avoiding Z. Thus $C_1, \ldots, C_{k-1}$ are distinct. For each such constraint set C_j let χ_j consist of the constraint $p(a^*_{i_j} | a^*_{i_{j+1}}) = 1$ for some distinguished assignments $a^*_{i_j}, a^*_{i_{j+1}}$ to $A_{i_j}, A_{i_{j+1}}$ respectively. Moreover add the constraint $p(a^*_{i_1}) = 1/2$ to χ_1, by writing χ_1 as $(p(a^*_{i_1}|a^*_{i_2}) - 1)(p(a^*_{i_1}) - 1/2) = 0$. (It is straightforward to see that each χ_j can be written in the form $f_j(z_j) = 0$.) Let all other constraints $(\chi_k, \ldots, \chi_m)$ be vacuous. The constraints $\chi_1, \ldots, \chi_m$ thus defined are clearly consistent, and constrain $C_1, \ldots, C_m$ respectively.

Note that by rewriting the constraints $\chi_1, \ldots, \chi_{k-1}$ and discarding the vacuous constraints $\chi_k, \ldots, \chi_m$, one can repose the optimisation problem as one involving constraint sets $C'_1, \ldots, C'_{k-1}$ where $C'_j = \{A_{i_j}, A_{i_{j+1}}\}$ for $j = 1, \ldots, k-1$. These constraint sets lead to a constraint graph $\mathcal{G}'$ in which the only edges are those between A_{i_j} and $A_{i_{j+1}}$ for $j = 1, \ldots, k-1$. By applying Theorem 4.1 to $\mathcal{G}'$, we see that $A_{i_j} \perp\!\!\!\perp_p \{A_{i_{j+2}}, \ldots, A_{i_k}\} \mid A_{i_{j+1}}$ for $j = 1, \ldots, k-2$, and (since none of $A_{i_1}, \ldots, A_{i_k}$ are in Z) $A_{i_1} \perp\!\!\!\perp_p Z \mid A_{i_k}$ and $A_{i_1} \perp\!\!\!\perp_p Z$. So for any $z@Z$,

$$\begin{aligned}
p(a^*_{i_1}|a^*_{i_k}z) &= p(a^*_{i_1}|a^*_{i_k}) \\
&= \sum_{a_{i_2},\ldots,a_{i_{k-1}}} p(a^*_{i_1}|a_{i_2}\ldots a_{i_{k-1}}a^*_{i_k})p(a_{i_2}|a_{i_3}\ldots a_{i_{k-1}}a^*_{i_k})\ldots \\
&\qquad \ldots p(a_{i_{k-2}}|a_{i_{k-1}}a^*_{i_k})p(a_{i_{k-1}}|a^*_{i_k})
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{a_{i_2},\ldots,a_{i_{k-1}}} p(a^*_{i_1}|a_{i_2})p(a_{i_2}|a_{i_3})\ldots \\
&\quad \ldots p(a_{i_{k-2}}|a_{i_{k-1}})p(a_{i_{k-1}}|a^*_{i_k}) \\
&= 1
\end{aligned}
$$

(the last step follows since $p(a_{i_j}|a^*_{i_{j+1}}) = 0$ if $a_{i_j} \neq a^*_{i_j}$). On the other hand, $p(a^*_{i_1}|z) = p(a^*_{i_1}) = 1/2 \neq 1 = p(a^*_{i_1}|a^*_{i_k}z)$, so $A_{i_1} \rightleftharpoons A_{i_k} \mid Z$, as required. ■

In sum, the constraint graph $\mathcal{G}$ offers a practical representation of the independencies satisfied by the maximum entropy function on account of the structure of the constraint sets.

Let z denote the parameter matrix with rows z_i, for $i = 1, \ldots, m$. Then $(\mathcal{G}, z)$ is called a *Markov network* with respect to the factorisation of Equation 4. Having worked out the values of the constant multipliers $\mu, \lambda_1, \ldots, \lambda_m$ in Equation 4 one can recast the entropy maximisation problem as follows. Given z, one can determine x from the factorisation, and hence the task of finding the x-parameters of the maximum entropy function can be reduced to that of finding the z-parameters of the maximum entropy function. While there were $(\prod_{i=1}^n K_i) - 1$ free x-parameters, these are now determined by $\sum_{i=1}^m (\prod_{A_j \in C_i} K_j) - 1$ free z-parameters. Note that one would expect the number of values K_j that variable A_j can take to be independent of the number of variables n and subject to practical limits. Suppose then that some constant K provides an upper bound for the K_j. At the end of §3 I suggested that the sizes $|C_i|$ of the constraint sets would also be subject to practical limits: suppose that the $|C_i|$ are bounded above by a constant L. Then there are at most $m(K^L - 1)$ free z-parameters. Thus if the number of constraints m increases linearly with n then so does the number of required z-parameters — a dramatic reduction from the number of x-parameters (bounded above by $K^n - 1$) required under the original formulation of the problem.[11]

While the Markov network formulation offers the possibility of a reduction in the complexity of entropy maximisation, it leaves us with two tasks: (i) to find the values of the multipliers in the factorisation, and (ii) to find the values of the z-parameters which yield maximum entropy. Neither of these tasks are straightforward in general: (i) the multipliers must be determined from a system of $(\prod_{i=1}^n K_i)$ equations (one factorisation for each $v@V$), and (ii) the z-parameters must be determined either from the same large system of equations or numerically from an analogue of the large summation expression for entropy, Equation 2.

[11] In fact, the x-parameters are determined by their marginals on the cliques (maximal complete subgraphs) of $\mathcal{G}$ (see [Lauritzen 1996], page 40). There are at most n cliques, so if clique-size and the K_j are bounded above, then the x-parameters are determined by a number of parameters that is *at worst* linear in n.

It is somewhat easier, in fact, to move to a second reparameterisation. Having reduced the complexity of the problem by exploiting independencies, we shall move from a Markov network parameterisation to a Bayesian network parameterisation. This will allow some simplification of the above two tasks and will leave us with a practical representation of the agent's belief function to which standard algorithms for updating can more easily be applied.

5 From Markov to Bayesian Network

An undirected graph is *triangulated* if for every cycle involving four or more vertices there is an edge in the graph between two vertices that are non-adjacent in the cycle. The first step towards a Bayesian network representation of the maximum entropy probability function is to construct a triangulated graph $\mathcal{G}^T$ from the constraint graph $\mathcal{G}$. Of course this move is trivial when, as is often the case, the constraint graph $\mathcal{G}$ is already triangulated. Figure 1, for example, is already triangulated. If $\mathcal{G}$ is not already triangulated, one of a number of standard triangulation algorithms can be applied to construct $\mathcal{G}^T$.[12]

Next, re-order the variables in V according to *maximum cardinality search* with respect to $\mathcal{G}^T$: choose an arbitrary vertex as A_1; at each step select the vertex which is adjacent to the largest number of previously numbered vertices, breaking ties arbitrarily. Let $D_1, \ldots, D_l$ be the cliques of $\mathcal{G}^T$, ordered according to highest labelled vertex. Let $E_j = D_j \cap (\bigcup_{i=1}^{j-1} D_i)$ and $F_j = D_j \backslash E_j$, for $j = 1, \ldots, l$.

In our example involving Figure 1, $A_1, \ldots, A_5$ are already ordered according to a maximum cardinality search,

$$D_1 = \{A_1, A_2\}, D_2 = \{A_2, A_3, A_4\}, D_3 = \{A_3, A_5\},$$

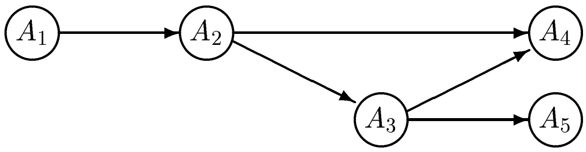
$$E_1 = \emptyset, E_2 = \{A_2\}, E_3 = \{A_3\},$$
$$F_1 = \{A_1, A_2\}, F_2 = \{A_3, A_4\}, F_3 = \{A_5\}.$$

Finally, construct an acyclic *directed constraint graph* $\mathcal{H}$ as follows. Take variables in V as vertices. Step 1: add an arrow from each vertex in E_j to each vertex in F_j, for $j = 1, \ldots, l$. Step 2: add further arrows to ensure that there is an arrow between each pair of vertices in $D_j, j = 1, \ldots, l$, taking care that no cycles are introduced (there is always some orientation of an added arrow which will not yield a cycle). In our example, an induced directed constraint graph $\mathcal{H}$ is depicted in Figure 2; the arrow from A_3 to A_4 was the only arrow added in step 2.

For disjoint $X, Y, Z \subseteq V$, Z *D-separates* X from Y in a directed acyclic graph if on each path between X and Y there is a structure $\longrightarrow A_i \longrightarrow$ or $\longleftarrow A_i \longrightarrow$ such that $A_i \in Z$, or a structure $\longrightarrow A_i \longleftarrow$ such that neither A_i nor its descendants are in Z.

[12]See for example [Neapolitan 1990] §3.2.3 and [Cowell et al. 1999] §4.4.1.

Figure 2: Example directed constraint graph



Theorem 5.1 If Z D-separates X from Y in the directed constraint graph $\mathcal{H}$ then $X \perp\!\!\!\perp_p Y \mid Z$ for any p satisfying the constraints which maximises entropy.

Proof: Since $\mathcal{G}^T$ is triangulated, the ordering yielded by maximum cardinality search is a *perfect* ordering (for each vertex, the set of its adjacent predecessors is complete in the graph).[13] Because the cliques are ordered according to highest labelled vertex where the vertices have a perfect ordering, the clique order has the *running intersection property* (for each clique, its intersection with the union of its predecessors is contained in one of its predecessors).[14] Now p factorises according to the cliques of $\mathcal{G}^T$, since it factorises according to $C_1, \ldots, C_m$ and these sets are complete in $\mathcal{G}^T$ and so are subsets of its cliques. These three facts imply that $p(v) = \prod_{i=1}^{l} p(f_i^v|e_i^v)$ for each $v@V$, where f_i^v, e_i^v are the assignments to F_i, E_i respectively which are consistent with v.[15]

Take an arbitrary component $p(f_i^v|e_i^v)$ of this factorisation. Each member of E_i is a parent (in $\mathcal{H}$) of each member of F_i and the members of F_i form a complete subgraph of $\mathcal{H}$ so we can write $F_i = \{A_{i_1}, \ldots, A_{i_k}\}$ where the parents of A_{i_j} are $Par_{i_j} =_{df} E_i \cup \{A_{i_1}, \ldots, A_{i_{j-1}}\}$. Hence,

$$\begin{aligned} p(f_i^v|e_i^v) &= p(a_{i_1}^v \ldots a_{i_k}^v|e_i^v) \\ &= \prod_{j=1}^{k} p(a_{i_j}^v|e_i^v a_{i_1}^v \ldots a_{i_{j-1}}^v) \\ &= \prod_{j=1}^{k} p(a_{i_j}^v|par_{i_j}^v) \end{aligned}$$

where $a_{i_j}^v$ and $par_{i_j}^v$ are the assignments to A_{i_j}, Par_{i_j} respectively that are consistent with v. Furthermore, each variable A_i occurs in precisely one F_j, so

$$p(v) = \prod_{i=1}^{n} p(a_i^v|par_i^v) \tag{5}$$

for each $v \in V$. When Equation 5 holds, p is said to *factorise* with respect to $\mathcal{H}$. It follows by a theorem due to Verma and Pearl that

[13] [Neapolitan 1990] Theorem 3.2.
[14] [Neapolitan 1990] Theorem 3.1.
[15] [Neapolitan 1990] Theorem 7.4.

if Z D-separates X from Y in $\mathcal{H}$ then $X \perp\!\!\!\perp_p Y \mid Z$.[16] ■

In general the directed constraint graph $\mathcal{H}$ is not as comprehensive a representation of independencies as the undirected constraint graph $\mathcal{G}$. If $\mathcal{G}$ is not already triangulated then some probabilistic independencies satisfied by entropy maximiser p in virtue of the structure of the constraint sets will not be implied by the directed constraint graph $\mathcal{H}$. To see this note that if $\mathcal{G} \neq \mathcal{G}^T$ then there must be two variables A_i and A_j which are not directly connected in $\mathcal{G}$, and so which are separated by some (possibly empty) Z in $\mathcal{G}$, but which are directly connected in $\mathcal{G}^T$ and thus in $\mathcal{H}$, and which are therefore not D-separated by Z in $\mathcal{H}$. On the other hand if $\mathcal{G} = \mathcal{G}^T$ then we do have an analogue of Theorem 4.2:

Theorem 5.2 Suppose $\mathcal{G}$ is triangulated. If, for all $\chi_1, \ldots, \chi_m$ constraining $C_1, \ldots, C_m$ respectively, $X \perp\!\!\!\perp_p Y \mid Z$ where p is a function satisfying $\chi_1, \ldots, \chi_m$ that maximises entropy, then Z D-separates X from Y in $\mathcal{H}$.

Proof: To check whether Z D-separates X from Y in $\mathcal{H}$ it suffices to check whether Z separates X from Y in the undirected *moral graph* formed by restricting $\mathcal{H}$ to X, Y, Z and their ancestors, adding an edge between any two parents in this graph that are not already directly connected, and replacing all arrows by undirected edges.[17] But all parents of vertices in $\mathcal{H}$ are directly connected, so the moral graph is a subgraph of $\mathcal{G}^T = \mathcal{G}$. By Theorem 4.2 if $X \perp\!\!\!\perp_p Y \mid Z$ for all such p then Z separates X from Y in $\mathcal{G}$. Hence Z separates X from Y in any subgraph of $\mathcal{G}$ that contains X, Y and Z, and in particular in the moral graph, as required. ■

Given some set $U \subseteq V$ containing A_i and its parents according to $\mathcal{H}$, and $u@U$, define parameter $y_i^u = p(a_i^u | par_i^u)$, where a_i^u, par_i^u are the assignments to A_i, Par_i respectively that are consistent with u. Let y_i be the vector of parameters y_i^u as u varies on A_i and its parents, and let y be the matrix with the y_i as rows, $i = 1, \ldots, n$. In this notation Equation 5 corresponds to

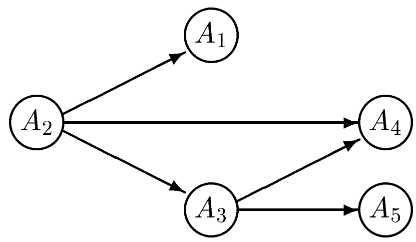
$$x^v = \prod_{i=1}^{n} y_i^v \tag{6}$$

for each $v@V$.

$(\mathcal{H}, y)$ is called a *Bayesian network*. Thanks to the factorisation of Equation 6, the task of finding x-parameters that maximise entropy can be reduced to that of finding the corresponding y-parameters. The number of free y-parameters required is determined by the cliques $D_1, \ldots, D_l$ in $\mathcal{H}$: there are $\sum_{i=1}^{l} (\prod_{A_j \in D_i} K_j) - 1$. Thus if clique-size

[16] See [Neapolitan 1990] Theorem 6.2.
[17] [Cowell et al. 1999] Corollary 5.11.

Figure 3: Alternative directed constraint graph



$|D_i|$ is bounded above by constant R and the number of values K_j bounded above by K, there are at most $n(K^R - 1)$ free y-parameters. If $\mathcal{G} \neq \mathcal{G}^T$ then the Bayesian network representation of p will require more parameters than the Markov network representation of §4. However the Bayesian network representation is more convenient for the following reasons.

First, there are no unknown multipliers in Equation 6. In contrast, in order to reconstruct the maximum entropy function from its Markov network representation via Equation 4, the values of constants $\mu, \lambda_1, \ldots, \lambda_m$ must be determined.

Second, the entropy equation can be reformulated in terms of the y-parameters as follows:

$$\begin{aligned}
H &= -\sum_{v@V} x^v \log x^v \\
&= -\sum_{v@V} \left(\prod_{j=1}^{n} y_j^v\right) \log \prod_{i=1}^{n} y_i^v \\
&= -\sum_{v@V} \left(\prod_{j=1}^{n} y_j^v\right) \sum_{i=1}^{n} \log y_i^v \\
&= -\sum_{i=1}^{n} \sum_{v@V} \left(\prod_{j=1}^{n} y_j^v\right) \log y_i^v \\
&= -\sum_{i=1}^{n} \sum_{v@Anc_i} \left(\prod_{A_j \in Anc_i} y_j^v\right) \log y_i^v
\end{aligned}$$

where $Anc_i \subseteq V$ consists of A_i and its ancestors in $\mathcal{H}$ (other terms cancel in the last step by additivity). In our example, Figure 2 induces an entropy equation of the form

$$H = -\sum_{v@A_1} y_1^v \log y_1^v - \sum_{v@\{A_1,A_2\}} y_1^v y_2^v \log y_2^v - \sum_{v@\{A_1,A_2,A_3\}} y_1^v y_2^v y_3^v \log y_3^v$$

$$- \sum_{v@\{A_1,A_2,A_3,A_4\}} y_1^v y_2^v y_3^v y_4^v \log y_4^v - \sum_{v@\{A_1,A_2,A_3,A_5\}} y_1^v y_2^v y_3^v y_5^v \log y_5^v$$

Note that roughly speaking there are fewest components in the sum of the entropy equation when the sets of ancestors Anc_i are smallest, and that when constructing $\mathcal{H}$, judicious use of maximum cardinality search and orientation of arrows can lead to a directed constraint graph with minimal ancestor sets. In our example, Figure 3 (where the vertices are labelled according to the original ordering, not that given by maximum cardinality search) is an alternative directed constraint graph, which leads to the following entropy equation:

$$H = -\sum_{v@A_2} y_2^v \log y_2^v - \sum_{v@\{A_1,A_2\}} y_1^v y_2^v \log y_1^v - \sum_{v@\{A_2,A_3\}} y_2^v y_3^v \log y_3^v$$
$$- \sum_{v@\{A_2,A_3,A_4\}} y_2^v y_3^v y_4^v \log y_4^v - \sum_{v@\{A_2,A_3,A_5\}} y_2^v y_3^v y_5^v \log y_5^v$$

This version of the entropy equation is more economical in the sense that the largest ancestor sets are smaller than those induced by Figure 2.

Having rewritten the entropy equation in terms of a y-parameterisation one can then use numerical techniques or Lagrange multiplier methods to find the values of the y-parameters that maximise H. If using the latter approach, note that there is an additivity constraint for each $i = 1, \ldots, n$ and each $u@Par_i$, of the form

$$\sum_{v@(\{A_i\}\cup Par_i), v\sim u} y_i^v = 1,$$

and each such constraint will require its own multiplier μ_i^u. Thus for assignment v to A_i and its parents, the partial derivative of the Lagrange equation takes the form

$$\frac{\partial H}{\partial y_i^v} + \mu_i^v + \sum_{i=1}^{m} \lambda_i \frac{\partial f_i}{\partial y_i^v} = 0$$

for

$$\frac{\partial H}{\partial y_i^v} = -\sum_{A_k : A_i \in Anc_k} \sum_{u@Anc_k, u\sim v} \left(\prod_{A_j \in Anc_k, j\neq i} y_j^u \right) [\log y_k^u + I_{k=i}]$$

where $I_{k=i} = 1$ if $k = i$ and 0 otherwise, and where as before $\lambda_i = 0$ for each inequality constraint χ_i which is not effective at y_i^v.

The third advantage of the Bayesian network parameterisation is this: the reparameterisation converts the general entropy maximisation problem into the special case problem of determining the parameters of a Bayesian network that maximise entropy; therefore we can apply existing techniques that have been developed for the special case to solve the general problem. Garside, Holmes, Markham and Rhodes have developed a number of efficient algorithms which determine the parameters of a Bayesian network that maximise entropy.

Their approach uses Lagrange multiplier methods on the original version of the entropy equation (Equation 2), subject to the restriction that the constraints must be linear. They have also developed specialised algorithms that deal with the cases in which the directed graph in the Bayesian network is a tree or inverted tree.[18] Schramm and Fronhöfer have investigated an alternative solution to the same problem, using an efficient system for maximising entropy that works by minimising cross entropy iteratively.[19]

Fourth, a Bayesian network is a good representation of an agent's belief function, given the uses such a function is normally put to, because Bayesian networks can be amenable to efficient calculations and updating.[20] There is now a large literature and set of computational tools for calculating marginals from a Bayesian network, and in particular conditional probabilities of the form $p(a_i|u)$, where $a_i @ A_i$ and $u @ U \subseteq (V \backslash \{A_i\})$.[21] Many such algorithms also implement Bayesian conditionalisation to update p on evidence u. Bayesian conditionalisation may be generalised to minimum cross entropy updating, which has similar justifications to those of the maximum entropy principle.[22] A minimum cross entropy update of x is a parameter vector x' which minimises

$$d(x, x') = \sum_{v @ V} x'^v \log \frac{x'^v}{x^v}$$

subject to any new constraints. By converting this to our y-parameterisation, it is not hard to see that the Bayesian network representation of p' will be the same as the Bayesian network representation of p on all variables except those in the new constraint sets and their predecessors under an ancestral ordering. Numerical methods or Lagrange multiplier methods can then be used with respect to the y-parameter formulation, in order to identify the new Bayesian network representation.[23]

6 Independence and Causal Constraints

Thus far we have been concerned only with the structure of the constraint sets, not with the nature of the constraints themselves. In this section we shall see that certain types of constraints — namely independence constraints and causal constraints — can be treated

[18] [Rhodes & Garside 1995], [Garside & Rhodes 1996], [Holmes & Rhodes 1998], [Rhodes & Garside 1998], [Garside et al. 1998], [Markham & Rhodes 1999], [Holmes et al. 1999], [Holmes 1999], [Garside et al. 2000].

[19] [Schramm & Fronhöfer 2002].

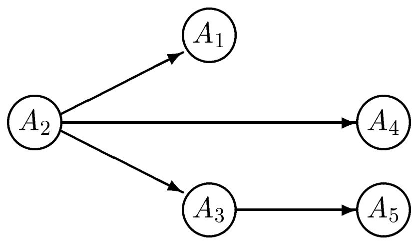
[20] Note that computational efficiency often depends on the structure of the directed graphs in Bayesian networks — very highly connected graphs are often not amenable to fast marginalisation or updating.

[21] See for example Part 1 of [Jordan 1998] and Chapter 6 of [Cowell et al. 1999].

[22] [Williams 1980].

[23] See [Williamson 2002].

Figure 4: Conditional independence constraint incorporated



especially easily, thanks to our Bayesian network representation of the maximum entropy function.

Many conditional independence constraints can be dealt with almost trivially. Suppose that in our example constraint χ_2 on constraint set $\{A_2, A_3, A_4\}$ says that A_3 and A_4 are probabilistically independent conditional on A_2, $A_3 \perp\!\!\!\perp_p A_4 \mid A_2$. Given the Bayesian network representation of Figure 3, this implies that for all assignments $u, v@\{A_2, A_3, A_4\}$ that agree on A_2 and A_4, $y_4^u = y_4^v$. This will hold just when p factorises according to Figure 4. Thus χ_2 can be satisfied just by manipulating $\mathcal{H}$, and changing the parent sets in the y-parameterisation accordingly — the constraint can then be ignored in the next step of the entropy maximisation process, namely that of finding the values of the new y-parameters that maximise entropy.

Note that it may not always be possible to adjust $\mathcal{H}$ to incorporate independence constraints. If, for example, χ_2 constrains p to satisfy $A_3 \perp\!\!\!\perp_p A_4 \mid A_2$ *and* $A_2 \perp\!\!\!\perp_p A_4 \mid A_3$, then there is no permutation of arrows amongst A_2, A_3, A_4 which will imply both these constraints and no others. The best one can do in this case is incorporate one of the constraints into the graph, and retain the other constraint in the optimisation of the corresponding y-parameters. Note also that if one can only represent a constraint by reversing one or more arrows in $\mathcal{H}$, one must check that the new graph is valid in the sense that there is a maximum cardinality search ordering of the vertices which would yield arrow directions compatible with the new arrow directions.

Causal constraints also have special consequences for entropy maximisation, as we shall see in the remainder of this section.

Judea Pearl articulated two problems with entropy maximisation in his pioneering book on Bayesian networks:

> computational techniques for finding a maximum-entropy [ME] distribution [Cheeseman 1983] are usually intractable, and the resulting distribution is often at odds with our perception of causation.[24]

The first problem — that of computational intractability — may be ameliorated by the methods put forward in the previous sections of

[24] [Pearl 1988] 463.

this paper. The second problem is that it is counterintuitive that adding an effect variable can lead to a change in the marginal distribution over the original variables:

> For example, if we first find an ME distribution for a set of n variables $X_1, \ldots, X_n$ and then add one of their consequences, Y, we find that the ME distribution $P(x_1, \ldots, x_n, y)$ constrained by the conditional probability $P(y|x_1, \ldots, x_n)$ changes the marginal distribution of the X variables ... and introduces new dependencies among them. This is at variance with the common conception of causation, whereby hypothesizing the existence of unobserved future events is presumed to leave unaltered our beliefs about past and present events. This phenomenon was communicated to me by Norm Dalkey and is discussed in [Hunter 1989].[25]

This problem is exemplified in 'Pearl's puzzle', which Daniel Hunter describes as follows.

> The puzzle is this: Suppose that you are told that three individuals, Albert, Bill and Clyde, have been invited to a party. You know nothing about the propensity of any of these individuals to go to the party nor about any possible correlations among their actions. Using the obvious abbreviations, consider the eight-point space consisting of the events $ABC, AB\bar{C}, A\bar{B}C$, etc. (conjunction of events is indicated by concatenation). With no constraints whatsoever on this space, MAXENT yields equal probabilities for the elements of this space. Thus $Prob(A) = Prob(B) = 0.5$ and $Prob(AB) = 0.25$, so A and B are independent. It is reasonable that A and B turn out to be independent, since there is no information that would cause one to revise one's probability for A upon learning what B does. However, suppose that the following information is presented: Clyde will call the host before the party to find out whether Al or Bill or both have accepted the invitation, and his decision to go to the party will be based on what he learns. Al and Bill, however, will have no information about whether or not Clyde will go to the party. Suppose, further, that we are told the probability that Clyde will go conditional on each combination of Al and Bill's going or not going. ...
>
> When MAXENT is given these constraints ... A and B are no longer independent! But this seems wrong: the information about Clyde should not make A's and B's actions dependent.[26]

[25] [Pearl 1988] 463-464.

[26] [Hunter 1989] 91.

To start with, when there are no constraints, the undirected constraint graph on A, B, C has no edges so by Theorem 4.1 the maximum entropy function yields all variables probabilistically independent. However when the probability distribution of C conditional on A and B is added as a constraint, the undirected constraint graph on A, B, C has an edge between each pair of variables. Thus by Theorem 4.2 there is some conditional probability distribution which renders A and B probabilistically dependent for the maximum entropy function. This dependence does indeed seem counterintuitive here.

The difficulty is that while we have taken into account the probability distribution of C conditional on A and B as a constraint on maximising entropy, we have ignored the further fact that A and B are causes of C. The key question is: how does causal information constrain the entropy maximisation process?

Hunter's answer to this conundrum is that causal statements are counterfactual conditionals and that the constraint in this example should be thought of as a set of probabilities of counterfactual conditionals rather than as a conditional probability distribution. Under Hunter's analysis of counterfactuals and probabilities of counterfactuals, a reconstruction of the above example retains the probabilistic independence of A and B when the constraint is added.

Hunter's response is in my opinion unconvincing, for two reasons. First, the counterfactual conception of causal relations adopted by Hunter is problematic. As Hunter himself acknowledges, his possible-worlds account of counterfactuals is rather simplistic.[27] More importantly though, the connection between causal relations and counterfactuals that Hunter adopts is implausible. Hunter says,

> the suggestion is that the relations between Al's and Bill's actions on the one hand and Clyde's on the other are expressible as counterfactual conditionals, that there is a certain probability that if Al and Bill *were* to go to the party, then Clyde *would not* go, and so on. The information to MAXENT should be probabilities of counterfactuals rather than conditional probabilities.[28]

This type of information is written in Hunter's notation using statements of the form $Prob(AB \Box\!\!\rightarrow C) = 0.1$. But such a statement expresses uncertainty about a counterfactual connection: the probability that Clyde would go were Al and Bill to go is 0.1. It does not express what we require, namely certain knowledge about a chancy causal connection, which would be better represented by $AB \Box\!\!\rightarrow (Prob(C) = 0.1)$: if Al and Bill were to go then Clyde would go with probability 0.1. In Pearl's puzzle we are told the exact causal relationships between A, B and C, and Hunter misrepresents these

[27] [Hunter 1989] 95.
[28] [Hunter 1989] 95.

as uncertain relationships. Moreover, correcting Hunter's representation of the causal connections seems unlikely to resolve Pearl's puzzle. In fact depending on how probability is interpreted one can even argue that $AB\Box\!\!\rightarrow (Prob(C) = 0.1)$ if and only if $Prob(C|AB) = 0.1$. For instance, under the Bayesian interpretation of probability $Prob(C|AB) = 0.1$ can be taken to mean that the agent in question would award betting quotient 0.1 to C were AB to occur; under the propensity interpretation it can be taken to mean that AB events have a (counterfactual) propensity to produce C events with probability 0.1. If this equivalence holds then Pearl's puzzle must still obtain, despite the counterfactual analysis.[29]

The second difficulty with Hunter's analysis is that while it resolves Pearl's puzzle, it fails to resolve a minor modification of Pearl's puzzle. In the original puzzle we are provided with the probability distribution of C conditional on A and B. Suppose instead we are provided with the distribution of C conditional on A, and the distribution of C conditional on B. In this case, the undirected constraint graph contains an edge between A and C and an edge between B and C; thus while $A \perp\!\!\!\perp_p B \mid C$ for maximum entropy p, there must be constraints with respect to which p will render A and B unconditionally dependent; this yields a puzzle analogous to that of the original problem. However Hunter's counterfactual reconstruction fails to eliminate the dependence of A and B in this modified puzzle.[30] In defence Hunter argues that his counterfactual analysis warrants the counterintuitive conclusion in the case of the modified puzzle, because according to his analysis situations in which A and B are positively correlated are more probable than situations in which A and B are negatively correlated. However, in the light of the above doubts about Hunter's analysis I suggest that intuition should prevail and that this new puzzle still needs resolving.

In fact I think that Pearl's puzzle and its modification can be resolved without having to appeal to a counterfactual analysis, any formulation of which is likely to be contentious. The resolution that I propose depends on making explicit the way in which qualitative causal relationships constrain entropy maximisation. Having made this constraint explicit, we shall see that it leads to a general framework for maximising entropy subject to causal knowledge. Finally, at the end of this section we shall see that the framework can be applied to resolve both Pearl's puzzle and its modification.

In [Williamson 2001] I suggested that causality satisfies a fundamental asymmetry, which can be elucidated with the help of the following example. Suppose an agent is concerned with two variables L and B signifying lung cancer and bronchitis respectively. Initially

[29] The relationship between causality and counterfactuals is in fact much more subtle than indicated here — see [Lewis 1973] — and many believe that there is no close relationship on account of these difficulties — see chapters 12-14 of [Sosa & Tooley 1993].

[30] [Hunter 1989] 101-104.

Figure 5: Smoking, lung cancer and bronchitis.

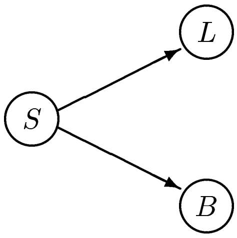
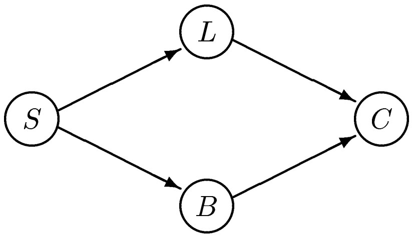


Figure 6: Smoking, lung cancer, bronchitis and chest pains.



she knows of no causal relationships between these variables, but she may have other background knowledge which leads her to adopt prior probability function p_1. Then the agent learns that smoking S causes each of lung cancer and bronchitis, which can be represented by a directed graph, Figure 5. One can argue that learning of the existence of common cause S should impact on her degrees of belief concerning L and B, making them more dependent. The reasoning is as follows: if an individual has bronchitis, then this may be because he is a smoker, and smoking may also have caused lung cancer, so the agent should believe the individual has lung cancer given bronchitis to a greater extent than before — the two variables become dependent (or more dependent if dependent already). Thus p_2, the new probability function determined with respect to her current knowledge (which includes the causal knowledge) might be expected to differ from p_1 over the original domain $\{L, B\}$.

Next the agent learns that both lung cancer and bronchitis cause chest pains C, giving the causal graph of Figure 6. But in this case one can *not* argue that L and B should be rendered more dependent. If an individual has bronchitis then he may well have chest pains, but this does not render lung cancer any more probable because there is already a perfectly good explanation for any chest pains. One cannot reason via a common effect in the same way that one can via a common cause, since learning of the existence of a common effect is irrelevant to an agent's current degrees of belief. Thus the new probability function p_3 ought to agree with p_2 on the domain of p_2, $\{S, L, B\}$.

This central asymmetry of causality can be explicated by what

I call the *causal irrelevance* condition. This says roughly that if an agent has initial belief function p_U on domain U and then learns of the existence of new variables which are not causes of any of the variables in U, then the restriction to U of her new belief function p_V on $V \supseteq U$ should agree with p_U on U, written $p_{V \restriction U} = p_U$. This condition can be rendered precise as follows.

Suppose that entropy is to be maximised subject to the causal knowledge represented by directed acyclic causal graph $\mathcal{C}$ on V, as well as the quantitative constraints $\chi = \{\chi_1, \ldots, \chi_m\}$ that we have considered in previous sections. Let $p_{\mathcal{C},\chi}$ denote the probability function that an agent adopts given her knowledge, $\mathcal{C}$ and χ. Given $U \subseteq V$ let $\mathcal{C}_U$ be the graph on U induced by $\mathcal{C}$ (the vertices are variables in U and the arrows correspond to those arrows in $\mathcal{C}$ between variables in U). We shall say that $\mathcal{C}$ is *irrelevant* to U with respect to χ if $p_{\mathcal{C},\chi \restriction U} = p_{\mathcal{C}_U,\chi \restriction U}$, i.e. the information in $\mathcal{C}$ that is not in $\mathcal{C}_U$ has no bearing on rational belief over U. A set of variables $U \subseteq V$ is a *ancestral* with respect to $\mathcal{C}$, or $\mathcal{C}$*-ancestral*, if it is closed under parents as determined by $\mathcal{C}$: $A_i \in U \Rightarrow Par_i^{\mathcal{C}} \subseteq U$. The causal irrelevance condition then says:

Causal Irrelevance If U is $\mathcal{C}$-ancestral and $\chi = \chi_U$ involves variables only in U then $\mathcal{C}$ is irrelevant to U with respect to χ_U, i.e. $p_{\mathcal{C},\chi_U \restriction U} = p_{\mathcal{C}_U,\chi_U}$.

My claim is that the causal irrelevance principle captures a key way in which causal knowledge constrains rational belief. Thus it is not enough to maximise entropy subject to quantitative constraints χ: one ought to take qualitative causal knowledge $\mathcal{C}$ into account too, by ensuring that causal irrelevance is satisfied. We shall see shortly how the causal irrelevance principle impinges on entropy maximisation.

First some further explanation of the causal irrelevance principle itself. The requirement that U is ancestral with respect to $\mathcal{C}$ is just the requirement that $V \backslash U$ must not contain any causes of variables in U. In the trivial case in which U is a singleton, $\mathcal{C}_U$ contains no causal information and we set $p_{\mathcal{C}_U,\chi_U} = p_{\chi_U}$, which may be found by maximising entropy subject to χ_U. In general the constraints in χ may involve all variables in V, not just those in U, in which case we can set χ_U to be the subset of those constraints in χ which only involve variables in U, $\chi_U = \{\chi_i : C_i \subseteq U, 1 \leq i \leq m\}$, and apply causal irrelevance to this restricted set of constraints. In this more general case causal irrelevance will tell that $p_{\mathcal{C},\chi_U \restriction U} = p_{\mathcal{C}_U,\chi_U}$, but nothing about $p_{\mathcal{C},\chi}$ or $p_{\mathcal{C},\chi \restriction U}$. The qualification that χ_U only involves variables in U is important, because knowledge of the existence of an effect together with probabilistic information about the effect itself can provide reason to change the probability distribution over its causes. If an agent with causal knowledge as in Figure 5 learns that lung cancer and bronchitis cause chest pains (Figure 6), *and she learns that the individual in question does have chest pains*, $p(c) = 1$, then arguably her degree of belief that the individual has lung cancer ought to be raised, and so too her degree of belief that he has bronchitis and her degree of belief that he is a smoker. Thus learning of non-causes and their probabilities can provide evidence to change current beliefs.

However, in some situations the probabilistic information about the non-causes $V \backslash U$ does not warrant any change in current beliefs, even when taken in conjunction with new causal knowledge: the new information is irrelevant to beliefs on the current domain U. More precisely we shall say χ is *irrelevant* to U with respect to $\mathcal{C}$ if $p_{\mathcal{C},\chi}\restriction_U = p_{\mathcal{C},\chi_U}\restriction_U$. For example the constraint $p(c|l \wedge b) = 0.9$ is intuitively irrelevant to $U = \{S, L, B\}$ with respect to $\mathcal{C}$ of Figure 6: learning of probabilities of effects conditional on their causes intuitively should not change degrees of belief over the causes. This intuition generalises as follows:

Probabilistic Irrelevance Suppose χ contains constraints χ_U on U together with constraints of the form $p(s|t) = w$ where $s@S \subseteq V \backslash U$, $t@T \subseteq V$, $w \in [0,1]$. If χ is *compatible* with $p_{\mathcal{C},\chi_U}\restriction_U$ in the sense that there is some function p satisfying χ that extends $p_{\mathcal{C},\chi_U}\restriction_U$, and U is $\mathcal{C}$-ancestral, then χ is irrelevant to U with respect to $\mathcal{C}$, i.e. $p_{\mathcal{C},\chi}\restriction_U = p_{\mathcal{C},\chi_U}\restriction_U$.

The compatibility qualification is just a kind of consistency condition: clearly χ can not be irrelevant to U if χ rules out the degrees of belief expressed by $p_{\mathcal{C},\chi_U}\restriction_U$.[31] The qualification that U should be ancestral with respect to $\mathcal{C}$ is also required: suppose $\mathcal{C}$ is the graph of Figure 5, $U = \{L, B\}, \chi_U = \emptyset$ but $\chi = \{p(s|l) = 0.99\}$; an agent may plausibly render L and B more probabilistically dependent under $\mathcal{C}$ and χ than under $\mathcal{C}$ and χ_U.

If χ is irrelevant to $\mathcal{C}$-ancestral U with respect to $\mathcal{C}$ then causal irrelevance tells us that

$$p_{\mathcal{C},\chi}\restriction_U = p_{\mathcal{C},\chi_U}\restriction_U = p_{\mathcal{C}_U,\chi_U} \tag{7}$$

Thus $\mathcal{C}$ and χ are jointly irrelevant to U.

In our example, the agent's belief function with respect to Figure 6 and $\chi = \{p(c|l \wedge b) = 0.9\} \cup \chi_U$ should, when restricted to $U = \{S, L, B\}$, be the same as her original belief function with respect to Figure 5 and original constraints χ_U.

We see then that irrelevance relates an agent's belief function over the domain V to her belief function over ancestral subdomains U. This fixes the Bayesian network representation of the probability function $p_{\mathcal{C},\chi}$ via the following results.

Theorem 6.1

- Order the variables in V ancestrally with respect to $\mathcal{C}$ (i.e. for each parent A_j of each $A_i, j < i$).

[31] See §11 of [Williamson 2002] for more on compatibility.

- Construct directed acyclic graph $\mathcal{H}$ on V by including an arrow to a variable from each predecessor that occurs in the same constraint set: $A_j \longrightarrow A_i$ iff $j < i$ and $A_i, A_j \in C_k$ for some $k, 1 \leq k \leq m$.

If χ is irrelevant to $U = \{A_1, \ldots, A_i\}$ with respect to $\mathcal{C}$ for each $i = 1, \ldots, n$, then

- ▸ Z D-separates X from Y in $\mathcal{H} \Rightarrow X \perp\!\!\!\perp_p Y|Z$ for $p = p_{\mathcal{C},\chi}$.

Proof: By Corollary 3 of [Pearl 1988] it is enough to show that $A_i \perp\!\!\!\perp_p A_i, \ldots, A_{i-1} \mid Par_i^{\mathcal{H}}$ for each $i = 1, \ldots, n$. We shall show this by induction on i. In the base case $i = 1$ there is nothing to do. Assume the inductive hypothesis for case $i-1$ and consider case i. Let $U = \{A_1, \ldots, A_i\}$. Clearly $A_i \perp\!\!\!\perp_{p_{\mathcal{C},\chi}} A_i, \ldots, A_{i-1} \mid Par_i^{\mathcal{H}}$ iff $A_i \perp\!\!\!\perp_{p_{\mathcal{C},\chi}|U} A_i, \ldots, A_{i-1} \mid Par_i^{\mathcal{H}}$. Since U is ancestral and χ is irrelevant to U, we have as in Equation 7 that $p_{\mathcal{C},\chi}|U = p_{\mathcal{C}_U,\chi_U}$ so we need to show that $A_i \perp\!\!\!\perp_{p_{\mathcal{C}_U,\chi_U}} A_i, \ldots, A_{i-1} \mid Par_i^{\mathcal{H}}$. But this holds by Theorem 4.1 since $Par_i^{\mathcal{H}}$ is the set of variables in U that occur in the same constraints as A_i. ■

Hence under these conditions $(\mathcal{H}, y)$ forms a Bayesian network representation of $p_{\mathcal{C},\chi}$, where the y parameters are defined by $y_i^u = p(a_i^u|par_i^u)$ as in §5. We saw in §5 that in the absence of causal knowledge the y parameters are the parameters that maximise $H = \sum_{i=1}^n H_i$ where

$$H_i = -\sum_{v@Anc_i} \left(\prod_{A_j \in Anc_i} y_j^v \right) \log y_i^v.$$

However when we have causal knowledge the situation is simpler yet: we can determine the y_1 parameters by maximising H_1, then the y_2 parameters by maximising H_2 subject to the y_1 parameters having been fixed in the previous step, and so on:

Theorem 6.2 Suppose as in Theorem 6.1 that the A_i are ancestrally ordered, χ is irrelevant to each $U = \{A_1, \ldots, A_i\}$, and $\mathcal{H}$ contains just arrows to A_i from predecessors that occur in the same constraint set. Then $p_{\mathcal{C},\chi}$ is represented by the Bayesian network $(\mathcal{H}, y)$ where the y_i maximise H_i subject to the constraints in $\chi_{\{A_1,\ldots,A_i\}}$, for $i = 1, \ldots, n$.

Proof: We shall use induction on i again. For the base case $i = 1$, set $U = \{A_1\}$. Then we proceed as in Equation 7: $y_i^u = p_{\mathcal{C},\chi}(a_1^u) = p_{\mathcal{C}_U,\chi_U}(a_1^u) = p_{\chi_U}(a_1^u)$ since the causal knowledge is trivial in this case. This is found by maximising entropy H on domain U subject only to χ_U, which is just maximising H_1 subject to χ_U. Assume the inductive hypothesis for case $i-1$ and consider case i. Setting $U = \{A_1, \ldots, A_i\}$ we have that $y_i^u = p_{\mathcal{C},\chi}(a_i^u|par_i^u) = p_{\mathcal{C}_U,\chi_U}(a_i^u|par_i^u)$. We find the y_i^u

by maximising H on domain U, i.e. $\sum_{j=1}^{i} H_j$, subject to χ_U. Now the $y_j^u, j = 1, \ldots, i-1$, are fixed by the inductive hypothesis, and hence so are the $H_j, j = 1, \ldots, i-1$. This it suffices to maximise H_i with respect to parameters y_i and subject to χ_U. ■

Thus when there is causal knowledge and the quantitative knowledge is irrelevant to each $U = \{A_1, \ldots, A_i\}$, the general entropy maximisation task, which requires simultaneously finding the y parameters that maximise H, reduces to the simpler task of sequentially finding the y_i parameters that maximise H_i, for $i = 1, \ldots, n$. Clearly this can offer enormous efficiency savings, both for numerical optimisation techniques and Lagrange multiplier methods. In the Lagrange multiplier case partial derivatives are simpler and each partial derivative involves only one free parameter — in fact it is straightforward to derive an analogue of Equation 4:

$$y_i^v = e^{\frac{\mu_i^v}{\pi} - 1} \prod_{C_i \subseteq \{A_1, \ldots, A_i\}} e^{\frac{\lambda_i}{\pi} \frac{\partial f_i}{\partial y_i^v}},$$

where the constant $\pi = \sum_{u @ Anc_i, u \sim v} \prod_{A_j \in Anc_i, j \neq i} y_j^u$ is fixed by having determined y_j^v for $j < i$ earlier in the sequential maximisation.

There are some important special cases. Suppose that each variable A_i occurs only with its direct causes $Par_i^{\mathcal{C}}$ in the constraint sets. If χ is irrelevant to each $U = \{A_1, \ldots, A_i\}$ then the independence graph $\mathcal{H}$ is just $\mathcal{C}$, the causal graph, and $(\mathcal{C}, y)$ offers a Bayesian network representation of $p_{\mathcal{C},\chi}$.[32] Suppose for example that all constraints take the form of probabilities of a variable conditional on assignments to their parents. Then compatibility and hence the probabilistic irrelevance of these constraints is guaranteed. If *each* probability of the form $y_i^u = p(a_i^u | par_i^u)$ is given as a constraint then the probability function $(\mathcal{C}, y)$ is fully determined by the causal graph and the constraints and no work is required to maximise entropy.[33] If some of these parameters are given then sequential maximisation can be used to determine the others.[34]

In Pearl's puzzle we are given the causal information $A \longrightarrow C, B \longrightarrow C$, and the conditional probability distribution of C conditional on A and B. By probabilistic irrelevance this conditional probability distribution is irrelevant to $\{A, B\}$ with respect to the causal information. By causal irrelevance the causal knowledge that C is common effect of A and B is irrelevant to $\{A, B\}$ (with respect

[32] This analysis can be used to provide an account of when we can expect the *causal Markov condition* to hold. The causal Markov condition says that the independencies implied by a causal graph under D-separation must hold in virtue of the graph being constructed causally. See [Pearl 2000], [Williamson 2001] and [Williamson 2002b] for more on this condition.

[33] Thus the approach here generalises that of [Williamson 2001].

[34] This is essentially the context in which [Lukasiewicz 2000] advocated sequential entropy maximisation. The framework here clearly provides a justification for his approach.

to $\chi_{\{A,B\}} = \emptyset$). Our analysis now tells us that the agent's probability function over $\{A, B, C\}$ is represented by a Bayesian network $(\mathcal{C}, y)$, where $\mathcal{C}$ is the graph capturing the causal information and the y parameters consist of the given conditional distribution together with $p(a) = 1/2$ and $p(b) = 1/2$ found by sequential entropy maximisation. In particular this probability function agrees with that formed on domain $\{A, B\}$ under no constraints. Thus we do not have any puzzling counterintuitive change in degrees of belief.

Moreover, the same reasoning goes through in the modification of Pearl's puzzle. Here we are given the same causal knowledge but the distribution of C conditional on A and that of C conditional on B, not that of C conditional on A and B. We now have to use sequential maximisation to provide the distribution of C conditional on A and B as parameters for a Bayesian network representation, but probabilistic and causal irrelevance still rid us of any counterintuitive dependency between A and B.

7 Maximum Entropy and Probabilistic Logic

The approach developed above has an interesting application to the foundations of probabilistic logic, as we shall now see.

A finite propositional language $\mathcal{L}$ may be thought of as a domain V of n variables $A_1, \ldots, A_n$, each of which is two-valued with possible assignments a_i and $\neg a_i$, for $i = 1, \ldots, n$. The sentences $\mathcal{SL}$ of $\mathcal{L}$ are constructed in the usual manner from these assignments, by applying the connectives $\wedge, \vee, \rightarrow, \leftrightarrow$. The probabilistic sentences $\mathcal{PSL}$ are of the form $p(\theta) = r$, where $\theta \in \mathcal{SL}$ and $r \in [0, 1]$.

The goal of a probabilistic logic is to decide whether and how a probabilistic sentence follows from a set of probabilistic sentences. One attempt at a semantics is to say $p(\theta_1) = r_1, \ldots, p(\theta_m) = r_m \models p(\phi) = s$ iff $p(\phi) = s$ follows by deductive logic from the axioms of probability and the premises $p(\theta_1) = r_1, \ldots, p(\theta_m) = r_m$. This yields rather a weak logic though, in the sense that the premises need not imply $p(\phi) = s$ for any value of s, via the axioms of probability. The premises often *underdetermine* the probability of ϕ.

More interesting, then, is a probabilistic logic with the semantics: $p(\theta_1) = r_1, \ldots, p(\theta_m) = r_m \models p(\phi) = s$ iff $p(\phi) = s$ for a function p, from all those that satisfy the constraints imposed by the premises, that maximises entropy.[35] The constraints imposed by the premises are linear with respect to the x-parameterisation,[36] and so if $\theta_1, \ldots, \theta_m, \phi$ are consistent there is a unique maximum entropy p satisfying the constraints, and thus a unique value s for which $p(\theta_1) = r_1, \ldots, p(\theta_m) = r_m \models p(\phi) = s$.

[35][Nilsson 1986] appears to have been the first to advocate this solution to the problem of probability underdetermination in the context of probabilistic logic. He acknowledges in §5 that his own computational methods become impractical in large problems.

[36][Paris 1994] pp. 13-14.

To complete the logic we require a proof theory such that

$$p(\theta_1) = r_1, \ldots, p(\theta_m) = r_m \vdash p(\phi) = s$$

if and only if

$$p(\theta_1) = r_1, \ldots, p(\theta_m) = r_m \models p(\phi) = s.$$

Thus if we can apply the methods developed in preceding sections to develop a practical procedure for finding the value of $p(\phi)$ for maximum entropy p, given the constraints $p(\theta_1) = r_1, \ldots, p(\theta_m) = r_m$, we could then compare this value with s in order to decide the probabilistic consequence.

In this case the constraint sets C_i are the variables whose assignments occur in θ_i, for $i = 1, \ldots, m$. The procedure outlined in this paper may be used to construct a Bayesian network that represents the maximum entropy probability function p. Let V_ϕ be the variables whose assignments occur in ϕ. Note that $p(\phi) = \sum_{v @ V_\phi, v \models \phi} p(v)$. Thus by querying the Bayesian network to find these $p(v)$ one can determine the correct value for s. If few variables occur in each θ_i in comparison with n as n becomes large then the constraint sets will be small relative to n, the induced Bayesian network correspondingly sparse, and the querying for $p(v)$ correspondingly quick. We have, then, a fully general proof procedure for probabilistic logic which promises to be practical for a range of problems.[37]

8 Concluding Remarks

Merely given the sets of variables that feature in constraints imposed by background knowledge one can achieve significant reductions in the number of parameters required to specify a maximum entropy probability function. I have described two representations of the maximum entropy function — a Markov network and a Bayesian network representation. A Markov network offers a natural representation of the independencies determined by the constraint sets, while a Bayesian network captures many if not all of these independencies and moreover offers a useful factorisation in terms of conditional probabilities of variables given their parents.

The approach presented here is perhaps best viewed as an extension or enhancement of, rather than an alternative to, several of the other proposed solutions to the entropy maximisation problem. First, as noted earlier, simple numerical or Lagrange multiplier techniques can be very inefficient when directly applied to entropy maximisation in its standard parameterisation. However, by applying these

[37]For an in-depth discussion of the foundations of probabilistic logic see [Williamson 2002c]. [Lukasiewicz & Kern-Isberner 1999] discusses maximum entropy applied to probabilistic logic programming.

methods to the Bayesian network parameterisation, dramatic efficiency savings can be made.[38] Second, the approach presented here can be viewed as an extension of the entropy maximisation techniques of Garside, Holmes, Markham and Rhodes, and of Schramm and Fronhöfer, from Bayesian networks to the general case. Third, Cheeseman presented a method for exploiting the log-linear factorisation of the maximum entropy function to compute conditional probabilities more efficiently,[39] and the present approach takes this a step further by enabling all the machinery developed for calculating conditional probabilities from Bayesian networks to be directly applied to the maximum entropy function.[40]

I have argued that the reparameterisations are most useful when the constraint sets are small in comparison with n, as n grows. I suggested that small constraints sets are the norm, but I have made no attempt to analyse their ubiquity here. This is an empirical claim and some form of empirical testing of this claim, as well as an empirical analysis of the practical performance of the reparameterisation approach, would clearly be of interest as the next step in developing this research. It remains to be seen whether small constraint sets are predominant in applications of probabilistic logic, and in other specialised applications of maximum entropy reasoning such as communication theory[41] and statistical physics.[42] Note that natural language processing has recently become an important application domain for maximum entropy techniques.[43] In natural language processing one often considers constraints imposed by context on the meaning of terms. The number of context variables that can be considered in these constraints is not fixed in advance but is dictated by available computational power. By pursuing a Bayesian network parameterisation, one may therefore be able to increase the size of the contexts that can be used to ascertain meaning.

It would also be interesting to look more closely at implemen-

[38] Of course in certain situations the problem may be simple enough not to warrant reparameterisation. If there are a small number of linear constraints, for example, then Lagrange multiplier methods can be used quite efficiently by converting to unconstrained dual form and optimising with respect to the multipliers as parameters.

[39] [Cheeseman 1983].

[40] Note that while Goldman and Rivest also build on Cheeseman's approach and also construct graphs as part of the entropy maximisation process ([Goldman & Rivest 1986], [Goldman & Rivest 1988]), their resulting proposal is quite different from the one presented here: they restrict attention to the case in which constraints are marginal probabilities; they construct a hypergraph (a graph involving sets of variables as vertices) and suggest the collecting of new data, and thus new constraints, in order to make the hypergraph acyclic; when this is achieved, the maximum entropy function can be calculated from the marginal constraints.

[41] [Gray 1990].

[42] http://bayes.wustl.edu/

[43] [Berger et al. 1996], [Ratnaparkhi 1998], [Ratnaparkhi 1999], [Borthwick 1999], [Nigam et al. 1999], [Wu & Khudanpur 2000], [Charniak 2000], [Varea et al. 2001], [Mullen et al. 2001], [Osborne 2002].

tational issues. We have seen, for instance, that different maximal cardinality search orderings will yield different directed constraint graphs from the same undirected constraint graph, some of which lead to an entropy equation with fewer terms: is there a way of quickly identifying an optimal directed constraint graph? A variety of numerical techniques are used for maximising entropy with respect to the standard parameterisation:[44] how do they compare on the Bayesian network parameterisation? How do Lagrange multiplier methods fare on the Bayesian network parameterisation?

While several questions remain open, I hope to have shown here that the construction of a Bayesian network from constraints offers the prospect of efficient entropy maximisation.

References

[Berger et al. 1996] Adam Berger, Stephen Della Pietra & Vincent Della Pietra: 'A maximum entropy approach to natural language processing', Computational Linguistics 22(1), pages 39-71.

[Borthwick 1999] Andrew Borthwick: 'A Maximum Entropy Approach to Named Entity Recognition', PhD Thesis, New York University.

[Brown 1959] D. Brown: 'A note on approximations to discrete probability distributions', Information and Control 2, pages 386-392.

[Charniak 2000] Eugene Charniak: 'A Maximum-Entropy-Inspired Parser', in Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics.

[Cheeseman 1983] Peter Cheeseman: 'A method of computing generalised Bayesian probability values for expert systems', In Proceedings of the 6th International Joint Conference on Artificial Intelligence, pages 198-202.

[Corfield & Williamson 2001] David Corfield & Jon Williamson(eds.): 'Foundations of Bayesianism', Kluwer Applied Logic Series.

[Cowell et al. 1999] Robert G. Cowell, A. Philip Dawid, Steffen L. Lauritzen & David J. Spiegelhalter: 'Probabilistic networks and expert systems', Springer.

[Csiszár 1989] I. Csiszár: 'A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling', The Annals of Statistics 17(3), pages 1409-1413.

[44] [Brown 1959], [Darroch & Ratcliff 1972], [Csiszár 1989].

[Darroch & Ratcliff 1972] J. Darroch & D. Ratcliff: 'Generalized iterative scaling for log-linear models', Annals of Mathematical Statistics 43, pages 1470-1480.

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

[Garside et al. 1998] G.R. Garside, D.E. Holmes & P.C. Rhodes: 'Using Maximum Entropy to Estimate Missing Information in Tree-like Causal Networks', in Proceedings of the 7th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, La Sorbonne, Paris, France, pages 359-366.

[Garside et al. 2000] G.R. Garside, D.E. Holmes & P.C. Rhodes: 'Using Maximum Entropy to Estimate Missing Information in Tree-like Causal Networks', in B. Bouchon-Meunier, R.R. Yager & L.A. Zadeh(eds): 'Advances in Fuzzy Systems - Application and Theory', Vol 20, World Scientific, pages 174-184.

[Goldman & Rivest 1986] Sally A. Goldman & Ronald L. Rivest: 'A non-iterative maximum entropy algorithm', in Proceedings of the Second International Conference on Uncertainty in Artificial Intelligence, Elsevier, pages 133-148.

[Goldman & Rivest 1988] Sally A. Goldman & Ronald L. Rivest: 'Making maximum entropy constraints easier by adding extra constraints (extended abstract)', in G.J. Erickson & C.R. Smith(eds.): 'Maximum-entropy and Bayesian methods in science and engineering', Volume 2, Kluwer, pages 323-340.

[Gray 1990] Robert M. Gray: 'Entropy and information theory', Springer-Verlag.

[Halpern & Koller 1995] Joseph Y. Halpern & Daphne Koller: 'Representation dependence in probabilistic inference', Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI 95), pages 1853-1860.

[Holmes 1999] D.E. Holmes: 'Efficient Estimation of Missing Information in Multivalued Singly Connected Networks using Maximum Entropy', in W. von der Linden et al.(eds): 'Maximum Entropy and Bayesian Methods', Kluwer Academic Publishers, pages 289-300.

[Holmes & Rhodes 1998] D.E. Holmes & P.C. Rhodes: 'Reasoning with Incomplete Information in a Multivalued Multiway Causal Tree Using the Maximum Entropy Formalism', International Journal of Intelligent Systems 13, pages 841-858.

[Holmes et al. 1999] D.E. Holmes, P.C. Rhodes & G.R. Garside: 'Efficient Computation of Marginal Probabilities in Multivalued Causal Inverted Multiway Trees given Incomplete Information', International Journal of Intelligent Systems 12, pages 101-111.

[Hunter 1989] Daniel Hunter: 'Causality and maximum entropy updating', International Journal in Approximate Reasoning 3, pages 87-114.

[Jaynes 1957] E.T. Jaynes: 'Information theory and statistical mechanics', The Physical Review 106(4), pages 620-630.

[Jordan 1998] Michael I. Jordan(ed.): 'Learning in Graphical Models', Cambridge, Massachusetts: The M.I.T. Press 1999.

[Lauritzen 1996] Steffen L. Lauritzen: 'Graphical models', Clarendon Press.

[Lewis 1973] David K. Lewis: 'Causation', with postscripts in [Lewis 1986], pages 159-213.

[Lewis 1986] David K. Lewis: 'Philosophical papers volume II', Oxford University Press.

[Lukasiewicz 2000] Thomas Lukasiewicz: 'Credal networks under maximum entropy', in Craig Boutilier & Moisés Goldszmidt (eds.), Proceedings of the Sixteenth Conference in Uncertainty in Artificial Intelligence, Morgan Kaufmann, pages 363-370.

[Lukasiewicz & Kern-Isberner 1999] T. Lukasiewicz & G. Kern-Isberner: 'Probabilistic logic programming under maximum entropy', in Proceedings of the 5th European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty, volume 1638 of LNAI, Springer, pages 279-292.

[Markham & Rhodes 1999] M.J. Markham & P.C. Rhodes: 'Maximising Entropy to deduce an initial probability distribution for a Causal Network', International Journal of Uncertainty, Fuzzyness and Knowledge-Based Systems 7(1), pages 63-68.

[Mullen et al. 2001] Tony Mullen, Robert Malouf & Gertjan van Noord: 'Statistical parsing of Dutch using maximum entropy models with feature merging', in Proceedings of the 6th Natural Language Processing Pacific Rim Symposium.

[Neapolitan 1990] Richard E. Neapolitan: 'Probabilistic reasoning in expert systems: theory and algorithms', New York: Wiley.

[Nigam et al. 1999] Kamal Nigam, John Lafferty & Andrew McCallum: 'Using maximum entropy for text classification', in Proceedings of the IJCAI-99 Workshop on Machine Learning for Information Filtering.

[Nilsson 1986] Nils J. Nilsson: 'Probabilistic logic', Artificial Intelligence 28, pages 71-87.

[Osborne 2002] Miles Osborne: 'Using Maximum Entropy for Sentence Extraction', in Proceedings of the ACL 2002 Workshop on Automatic Summarization (including DUC 2002), Philadephia, Pennsylvania USA, July 11-13 2002.

[Paris 1994] J.B. Paris: 'The uncertain reasoner's companion', Cambridge: Cambridge University Press.

[Paris & Vencovská 1997] J.B. Paris & A. Vencovská: 'In defense of the maximum entropy inference process', International Journal of Automated Reasoning 17, pages 77-103.

[Paris & Vencovská 2001] J.B. Paris & A. Vencovská: 'Common sense and stochastic independence', in [Corfield & Williamson 2001], pages 203-240.

[Pearl 1988] Judea Pearl: 'Probabilistic reasoning in intelligent systems: networks of plausible inference', Morgan Kaufmann.

[Pearl 2000] Judea Pearl: 'Causality: models, reasoning, and inference', Cambridge University Press.

[Ratnaparkhi 1998] Adwait Ratnaparkhi: 'Maximum Entropy Models for Natural Language Ambiguity Resolution', PhD Thesis, University of Pennsylvania.

[Ratnaparkhi 1999] Adwait Ratnaparkhi: 'Learning to Parse Natural Language with Maximum Entropy Models', Machine Learning 34, pages 151-175.

[Rhodes & Garside 1995] P.C. Rhodes & G.R. Garside: 'Using maximum entropy to compute marginal probabilities in a causal binary tree need not take exponential time', in C. Froidevaux and J. Kohlas(eds.): Proceedings of ECSQARU'95: Symbolic and Quantitative Approaches to Reasoning and Uncertainty, Springer, pages 352-363.

[Rhodes & Garside 1998] P.C. Rhodes & G.R. Garside: 'Computing marginal probabilities in causal inverted binary trees given incomplete information', Knowledge-Based Systems 10, pages 213-224.

[Schramm & Fronhöfer 2002] Manfred Schramm & Bertram Fronhöfer: 'Completing incomplete Bayesian networks', Proceedings of the Workshop on Conditionals, Information and Inference, FernUniversität May 13-15 2002, pages 231-244.

[Sosa & Tooley 1993] Ernest Sosa & Michael Tooley(eds.): 'Causation', Oxford: Oxford University Press.

[Sundaram 1996] Rangarajan K Sundaram: 'A first course in optimisation theory', Cambridge: Cambridge University Press.

[Varea et al. 2001] Ismael Garcia Varea, Franz Josef Och, Hermann Ney & Francisco Casacuberta: 'Refined Lexicon Models for Statistical Machine Translation using a Maximum Entropy Approach', in Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics', Toulouse, France, July 2001, pages 204-211.

[Williams 1980] Peter M. Williams: 'Bayesian conditionalisation and the principle of minimum information', British Journal for the Philosophy of Science 31, pages 131-144.

[Williamson 2001] Jon Williamson: 'Foundations for Bayesian networks', in [Corfield & Williamson 2001], pages 75-115.

[Williamson 2002] Jon Williamson: 'Bayesianism and language change', Journal of Logic, Language and Information, to appear.

[Williamson 2002b] Jon Williamson: 'Learning causal relationships', Discussion Paper, LSE Centre for Natural and Social Sciences, available as report jw02d at www.kcl.ac.uk/philosophy.ai.

[Williamson 2002c] Jon Williamson: 'Probability logic', in Dov Gabbay, Ralph Johnson, Hans Juergen Ohlbach & John Woods (eds.): 'Handbook of the Logic of Argument and Inference: The Turn Toward the Practical', Elsevier, pages 393-419.

[Wu & Khudanpur 2000] Jun Wu & Sanjeev Khudanpur: 'Efficient Training Methods for Maximum Entropy Language Modeling', in Proceedings of the International Conference on Spoken Language Processing Vol. 3, Oct. 2000, Beijing, China, pages 114-117.