CHORDAL GRAPHS TO IDENTIFY GRAPHICAL MODELS SOLUTIONS OF MAXIMUM OF ENTROPY UNDER CONSTRAINTS ON MARGINALS

ALAIN FRANC*, MICHEL GOULARD†, AND NATHALIE PEYRARD‡

Abstract. We consider the problem of specifying the joint distribution of a collection of variables with maximum entropy when a set of marginals are fixed. One can easily derive that the structure of the solution joint distribution is that of a graphical model. The potential functions are then marginals at some power. We address the following question: under which conditions on the set of constraints, is it possible to fully identify the canonical exponents in the maximum entropy solution as functions of the problem structure? Literature related to this topic is somehow scattered in disciplines such as statistical mechanics, information theory, graph theory or inference in graphical models. In this article we gather and link results from these different fields. From this, we show that for a particular class of constraints set on marginal, the *chordal maximal coherent sets of constraints*, it is possible to derive the canonical exponents of the graphical model solution of the maximum entropy problem as the numbers of occurrence of separators in an associated joint tree. Conversely, we present sufficient conditions to ensure that a graphical model is solution of a maximum entropy problem.

Key words. maximum entropy, graphical models, chordal graphs.

AMS subject classifications. 05C, 60J, 68R10, 94A17

1. Introduction. Consider the problem of infering or specifying the joint distribution of a collection of variables, from partial information represented by empirical knowledge on some marginal distributions. Each marginal distribution involves only subsets (not necessarily disjoint) of the collection. One constructive approach to determine a joint distribution coherent with this information is the principle of maximum entropy under constraints ([10]). This method ensures that the extra amount of information embodied into the solution distribution, in addition to the one brought by the the constraints is minimal. For some particular set of constraints, the solution of the maximum entropy (ME) problem is well know and leads to classical distributions: uniform distributions are solution of the ME problem when no information is available; the Gibbs distributions maximise the entropy among distributions with same fixed mean; Gaussian distributions can be recovered as ME distributions for fixed mean and variance. In these three situations, the solution of ME under constraints can be expressed straightforwardly as a function of the constraints.

When constraints are specifications on some marginals of the distribution, if two fixed marginals involve two non disjoint subsets of random variables, then these two constraints imply conditional independence properties, in the Markovian sense, in the ME solution. One can easily derive that the structure of solution joint distribution is that of a graphical model ([23], this property of the ME solution holds actually for any set of linear constraints). More precisely, the joint distribution is a product of some marginals at a certain power. Exponents are called canonical exponents ([22]). In general the ME distribution is solution of a system intractable by hand, and identification of the canonical exponents is not possible by simple calculus.

*INRA, UMR 1202, Biodiversité, Gènes et Communautés - 69, route d'Arcachon - Pierroton - 33612 Cestas Cedex - France

†INRA, UMR 1201, DYNAFOR - Chemin de Borde Rouge, 31326 Castanet-Tolosan - France

‡INRA, UR 875, Unité de Biométrie et Intelligence Artificielle - Chemin de Borde Rouge, 31326 Castanet-Tolosan - France

In this paper we bring some insights on the following questions: for which particular structure of the set of specified marginals, is it possible to fully identify the canonical exponents of the ME solution? Under which conditions can a graphical model be derived as the solution of a ME problem? In very simple situations, the ME problem is easy to solve: if the marginals are specified on disjoint subsets of the variables, the distribution defined as the product of these marginals is the ME solution; if marginals are specified on pairs of variables only, and if these contraints induce no loop, then the ME solution is a graphical model on a tree, whose analytical expression is a classical result of graphical models ([19]): pairs marginals occur at power one and singleton marginals at power equal to the node degree minus one. How can we generalise these two results? We gather several well know results from maximum entropy methods ([10]), graph theory ([1]), graphical models ([16]) and variational methods for graphical models ([22]) to define a class of constraints sets on marginals, *chordal maximum sets of constraints*, for which we can derive the canonical exponents from the problem's structure. As opposed to the classical approach for solving ME problem, the main result of this work does not require Lagrange multipliers.

The paper is organised as follow. Basic notions on Probability and marginal distributions are recalled in Section 2. The maximum entropy problem is formalised in section 3 and its link with graphical models is developped in Section 4. Identification of the canonical exponents is illustrated on some toy examples (Section 5) to give the flavour of a key condition for analytical resolution of the ME problem: the existence of an elimination order of the variables. We present briefly some elements of graph theory to formalise this notion in Section 6 and present our main result for the familly of chordal maximum sets of constraints (Section 7). This papers ends by a discussion of link with methodologies for graphical models inference and variationnal approximations.

2. Coherent set of constraints on marginal distributions. Let us consider $\mathcal{V}$ a finite set of n points , with elements indexed on $\{1,\dots,n\}$. A random variable X_i is attached to each point $i \in \mathcal{V}$, taking value in Λ. Then, a state of $X = (X_1,\dots,X_n)$ is a vector $x = (x_1,\dots,x_n)$, with $x_i \in \Lambda$. The set of all possible states is called the state space, and is noted $\Omega = \Lambda^{|\mathcal{V}|}$. Ω_I is then the state space of the collection of random variables $X_I = \{X_i, i \in I\}$. If p is the probability distribution of X on Ω, we define the notation

$$p\,(x) = p\{X = x\}$$

A marginal distribution of p is the probability distribution of a subset of X induced by p. If $I \subset \mathcal{V}$ and $\bar{I} = \mathcal{V} \setminus I$, the marginal distribution p_I of X_I is defined as

$$p_I(x_I) = \sum_{x_{\bar{I}} \in \Omega_{\bar{I}}} p\,(x_I, x_{\bar{I}})$$

If p_I is given, then the marginal for any part J of I is uniquely specified as (with $K = I \setminus J$)

$$p_J(x_J) = \sum_{x_K} p_I(x_J, x_K) \tag{2.1}$$

If $\mathcal{M}$ is a set of parts $I \subset \mathcal{V}$, let us consider a set of specified marginal distributions on $\mathcal{M}$: $p_I = a_I, \forall I \in \mathcal{M}$. Marginal distributions of a set $\mathcal{M}$ are said *mutually coherent*

if for any intersection K of two elements I and J of $\mathcal{M}$, the marginal p_K derived from p_I is equal to the one derived from p_J:

$$\forall\, x_K \in \Lambda^{|K|}, \quad \sum_{x_{I\setminus K}} p_I(x_K, x_{I\setminus K}) = \sum_{x_{J\setminus K}} p_J(x_K, x_{J\setminus K}) = p_K(x_K)$$

Since when marginals on $\mathcal{M}$ are specified they are specified (by (2.1)) on any part of any element of $\mathcal{M}$, it is useful to consider the set

$$\mathcal{A} = \mathcal{M} \cup \{J \,:\, \exists\, I \in \mathcal{M} \,:\, J \subset I\}$$

The set $\mathcal{A}$ is a *complete set*, it is stable by intersection and inclusion.

The association of a complete set $\mathcal{A}$ of part of $\mathcal{V}$, and of mutually coherent marginal distributions $\{a_I, I \in \mathcal{A}\}$ is called a *coherent set of constraints* (CSC). It will be synthetically denoted by $(\mathcal{A}, a_I)$. For a CSC $(\mathcal{A}, a_I)$, a part $I \in \mathcal{A}$ is said *maximal* if there exists no element J in $\mathcal{A}$, such that $I \subset J$. The set of maximal elements of $\mathcal{A}$ is called the set of *generators* of $\mathcal{A}$ and will be denoted $\mathbb{G}(\mathcal{A})$.

A graph $\mathcal{G}_{CSC}$ can be associated to any CSC : vertices are points of $\mathcal{V}$ and two vertices i and j are linked by an edge if there exists $I \in \mathcal{A}$ such that $i \in I$ and $j \in I$. This graph will be referred to as the *constraint graph*, a classical notion of constraint processing ([5]).

3. Maximum entropy problem. The entropy of a probability distribution p is defined as

$$\mathcal{H}(p) = -\sum_x p\,(x) \operatorname{Log} p\,(x)$$

with convention that $p\,(x) \operatorname{Log}\, p\,(x) = 0$ whenever $p\,(x) = 0$. Finding the distribution with largest entropy under some constraints has been a classical way to either construct or interprete/justify classical probability distributions. The solution is the distribution satisfying the constraints and including minimal extra information, in addition to the one brought by the constraints: all the uncertainty permitted by the available information is maintained. When considering a set of random variables $X = (X_1, \ldots, X_n)$ with state space Ω, and a set of constraints taking the form of a specified set of marginals, the question is to find the simplest joint distribution (in the sense of its dependence structure) satisfying these constraints. Let $(\mathcal{A}, a_I)$ be a CSC . We can consider the set $\mathcal{Q}$ of distributions q on Ω which fulfill the constraints on each marginal a_I:

$$\mathcal{Q} = \{q \,:\, \forall\, I \in \mathcal{A}, \quad q_I = a_I\}$$

The solution (if it exists) of the maximum entropy (ME) problem under constraints $(\mathcal{A}, a_I)$ is then

$$q = \arg\max_{q \in \mathcal{Q}} \mathcal{H}(q)$$

Maximum entropy problem are classically solved using Lagrange multipliers ([10]), even if more efficient methods have been developped since ([8, 23] and reference therein). We will see in the next section that even if the explicit form of the solution is not always available, we can obtain some information on the structure of this solution, in terms of conditional independence assumptions, and factorisation of the solution joint distribution.

4. Maximum entropy and graphical models. A joint distribution p on Ω is said to be a *graphical model* ([16]) indexed on a set $\mathcal{B}$ of parts of $\mathcal{V}$ if there exists a set $\Psi = \{\psi_B\}_{B\in\mathcal{B}}$ of maps, called 'potentials', indexed by $\mathcal{B}$

$$\psi_B : \Omega_B \longrightarrow \mathbb{R}^+$$

such that p can be expressed in the following factored form:

$$p(x) = \prod_{B\in\mathcal{B}} \psi_B(x_B) \tag{4.1}$$

Note that a potential is not necessarily a probability distribution: in general ψ_B is not normalised and is not the marginal distribution of x_B. Classically, a graph $\mathcal{G}_{GM} = (\mathcal{V}, \mathcal{E})$ is associated to such a decomposition: the nodes are the points of $\mathcal{V}$ and an edge is drawn between two nodes i and j if there exists $B \in \mathcal{B}$ such that i and j are in B. If the set $\mathcal{B}$ forms a partition of $\mathcal{V}$, the factorisation property (4.1) implies that X_i and X_j are independent random variables if i and j are not included in the same part B of the partition. The graph of the model will be composed of $|\mathcal{B}|$ connected components. If intersections between elements of $\mathcal{B}$ are non empty, the factorisation of $p(x)$ implies conditional independence. For instance, if $\mathcal{V} = \{1, 2, 3\}$ and $\mathcal{B} = \{\{1, 2\}, \{2, 3\}\}$, in the corresponding graphical model, X_1 and X_3 are independent conditionally to x_2. The graph of the model is a 3-nodes line with 2 as central node, "separating" 1 from 2. More generally, if I, J and K are subsets of $\mathcal{V}$ such that any path in $\mathcal{G}_{GM}$ between a node in I and a node in J goes through K, then X_I is independent of X_J given x_K ([16]). We present now two properties on the structure of the solution of the ME under a given CSC , which specifies the relationship between graphical models and maximum of entropy when information available is on marginals.

LEMMA 4.1. *The joint probability distribution q with maximum of entropy compatible with a CSC $(\mathcal{A}, a_I)$ is a graphical model indexed by $\mathbb{G}(\mathcal{A})$:*

$$\forall\, x \in \Omega, \quad q\,(x) = \lambda \prod_{I\in\mathbb{G}(\mathcal{A})} \psi_I(x_I) \tag{4.2}$$

See the Appendix (Section 9) for the proof. Expression of the ME solution as a product of local functions, one for each constraint is not new and appears already in the seminal work of Jaynes ([10]). The explicitation of the link with graphical models can be found in [23].

For a general CSC , a major difficulty is to solve the system with contraints and to derive from (4.2) an analytic expression of the ME solution as a function of the constraints a_I. However, it is possible to specify further the decomposition of the solution of the ME problem:

LEMMA 4.2. *Let $(\mathcal{A}, a_I)$ be a CSC , any distribution on Ω which can be expressed as follows: .*

$$a(x) = \lambda \prod_{I\in\mathcal{A}} a_I(x_I)^{\beta_I} \tag{4.3}$$

with $\beta_I \in \mathbb{Z}$, and satisfies the CSC is of entropy larger that any distribution q on Ω satisfying the CSC :

$$\forall q \in \mathcal{Q}, \quad \mathcal{H}(q) \leq \mathcal{H}(a)$$

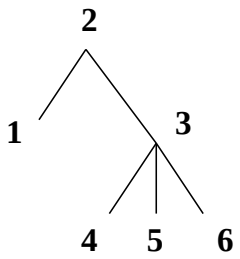

FIG. 5.1. *A CSC with tree structure:* $\mathbb{G}(\mathcal{A}) = \{\{1,2\},\{2,3\},\{3,4\},\{3,5\},\{3,6\}\}$.

The parameters β_i will be referred to as the canonical exponents, following the formalism of exponential families ([22]).

The proof is developped in the Appendix (Section 9). It differs from classical approaches for solving ME problem by the fact that it does not rely on the introduction of Lagrange multipliers.

Note that decomposition (4.2) runs on the set $\mathbb{G}(\mathcal{A})$ of generators of the CSC , while (4.3) runs on $\mathcal{A}$. Some exponents β_I may be equal to zero but we will see in section 7 that marginals on subsets other than the generators have non zero β_I.

From Lemma 4.2, we can state that if a distribution with decomposition (4.3) is constructed, which is coherent with the contraints, then, this distribution is the solution of the ME problem.

In the particular case where the generators of the CSC have empty intersections, resolution is tractable and $q(x) = \prod_{I\in\mathbb{G}(\mathcal{A}} a_I(x_I)$. Let us now consider a graphical model such that $\mathcal{G}_{GM} = \{\mathcal{V},\mathcal{E}\}$ is a tree. If a_I is the marginal distribution of the graphical model for I a pair of points linked by an edge in the tree, it can be shown ([19] and Section 5) that the distribution of the graphical model is given by

$$q\left(x\right) = \frac{\prod_{I\in\mathcal{E}} a_I(x_I)}{\prod_{i\in\mathcal{V}} a_i(x_i)^{d_i-1}} \tag{4.4}$$

where d_i is the degree of i, i.e. the number of points linked to i by an edge in $\mathcal{G}_{GM}$. From Lemma 4.2, this is the solution of the ME problem for the CSC $(\mathcal{A}, a_I)$ with $\mathcal{A} = \mathcal{V}\cup\mathcal{E}$. One can note that $\mathcal{G}_{GM} = \mathcal{G}_{CSC}$, meaning that the constraints set has a tree structure.

Under with condition on the structure of a CSC is it possible to fully specifiy the ME solution? In other words when is it possible to identify the exponents β_I? We explore this problem in the next sections of this article, starting with a set of simple examples to give a flavour of the important elements.

5. Examples. If the generators of a CSC $(\mathcal{A}, a_I)$ are of size 2 and $\mathcal{G}_{CSC}$ is a tree (see Fig. 5.1 for an example). We demonstrate here the optimality of distribution (4.4), starting from the decomposition (4.2). By integrating the normalisation constant in one of the potential functions, it is possible to write the maximum entropy solution as

$$p\left(x\right) = \prod_{i\sim j} \psi_{ij}(x_i, x_j)$$

where, for two nodes i and j of $\mathcal{G}_{CSC}$, notation $i\sim j$ means that there is an edge between i and j. Since $\mathcal{G}_{CSC}$ is a tree, there exists at least one vertex of degree equal

to one. Let us label it node 1, and its unique neighbour node 2. Then

$$p\left(x\right) = p\left(x_1 \,|\, x_2 \dots x_n\right) p\left(x_2 \dots x_n\right)$$

The conditional probability $p\left(x_1 \,|\, x_2 \dots x_n\right)$ can be written as a function of the potential ψ_{12} only as follows:

$$\begin{aligned} p\left(x_1 \,|\, x_2 \dots x_n\right) &= \frac{p\left(x_1 \dots x_n\right)}{\sum_{x_1' \in \Omega_1} p\left(x_1' \dots x_n\right)} \\ &= \frac{\psi_{12}(x_1, x_2) \prod_{i \sim j \,:\, i,j \neq 1} \psi_{ij}(x_i, x_j)}{\sum_{x_1' \in \Omega_1} \psi_{12}(x_1', x_2) \prod_{i \sim j \,:\, i,j \neq 1} \psi_{ij}(x_i, x_j)} \\ &= \frac{\psi_{12}(x_1, x_2)}{\sum_{x_1' \in \Omega_1} \psi_{12}(x_1', x_2)} \end{aligned}$$

Moreover, since p satisfies the marginals,

$$\begin{aligned} \frac{a_{12}(x_1, x_2)}{a_2(x_2)} &= \frac{\psi_{12}(x_1, x_2) \sum_{x_3' \dots x_n'} \left(\prod_{i \sim j \,:\, i,j \neq 1,2} \psi_{ij}(x_i', x_j') \right) \left(\prod_{i \neq 1} \psi(x_i', x_2) \right)}{\sum_{x_1' \in \Omega_1} \psi_{12}(x_1', x_2) \sum_{x_3' \dots x_n'} \left(\prod_{i \sim j \,:\, i,j \neq 1,2} \psi_{ij}(x_i', x_j') \right) \left(\prod_{i \neq 1} \psi(x_i', x_2) \right)} \\ &= \frac{\psi_{12}(x_1, x_2)}{\sum_{x_1' \in \Omega_1} \psi_{12}(x_1', x_2)} \end{aligned}$$

we obtain

$$p\left(x_1 \,|\, x_2 \dots x_n\right) = \frac{a_{12}(x_1, x_2)}{a_2(x_2)} = p\left(x_1 \,|\, x_2\right) \tag{5.1}$$

and we recover conditional independence property of graphical models: given x_2, the variable X_1 is independent of the other variables. Finally,

$$p\left(x\right) = \frac{a_{12}(x_1, x_2)}{a_2(x_2)} \, p\left(x_2 \dots x_n\right)$$

This procedure can be seen as a "deconditioning" of node 1. This procedure can be run through the whole tree $\mathcal{G}_{CSC}$ since if node 1 is remove, the resulting graph is still a tree, with at least one node of degree one, ... and so on. The treatment of a node i adds a term $\frac{a_{ij}(x_i, x_j)}{a_j(x_j)}$ in the expression of $p(x)$. The marginal a_j of node j is involved for the treatment of each node linked to j except one (the one used for the treatment of node j). If d_i is the degree of i by an edge in $\mathcal{G}_{CSC}$, n iterations of the procedure lead to

$$p\left(x\right) = \frac{\prod_{i \sim j} a_{ij}(x_i, x_j)}{\prod_{i \in \mathcal{V}} a_i(x_i)^{d_i - 1}}$$

For instance, the distribution of maximum of entropy with specified marginals $a_{12}, a_{23}, a_{34}, a_{35}, a_{36}$, (see Fig. 5.1) is

$$p = \frac{a_{12} a_{23} a_{34} a_{35} a_{36}}{a_1 a_3^2}$$

Note that the above demonstation does not explicitly exploit the fact that the edges of the constraint graph $\mathcal{G}_{\mathsf{CSC}}$ define conditional independence. This property is actually recovered via equation (5.1).

It is possible to identify the exponents in (4.3) for more complex structure of the graph $\mathcal{G}_{CSC}$. Let us now consider the two following CSC :

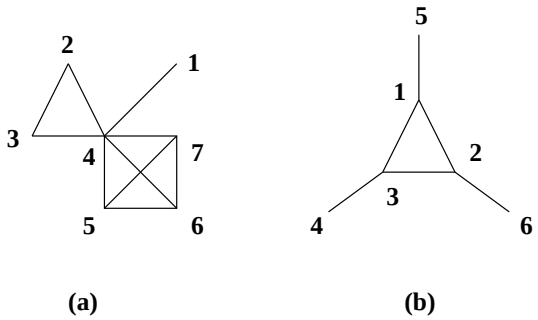

FIG. 5.2. *Two CSC with star structure: (a)* $\mathbb{G}(\mathcal{A}) = \{\{1,4\},\{2,3,4\},\{4,5,6,7\}\}$*, (b)* $\mathbb{G}(\mathcal{A}) = \{\{1,2,3\},\{1,5\},\{2,6\},\{3,4\}\}$.

(a) $\mathbb{G}(\mathcal{A}) = \{\{1,4\},\{2,3,4\},\{4,5,6,7\}\}$,
(b) $\mathbb{G}(\mathcal{A}) = \{\{1,2,3\},\{1,5\},\{2,6\},\{3,4\}\}$.

The corresponding $\mathcal{G}_{CSC}$ are represented on Fig. 5.2. Both present a star structure: an element $S \in \mathcal{A}$ has a central position in $\mathcal{G}_{CSC}$. This is respectively $\{4\}$ and $\{1,2,3\}$ on Fig. 5.2 (a) and (b). The set S has the following property that it separates the different generators of the CSC : any path in $\mathcal{G}_{CSC}$ from generator G_1 to generator G_2 goes through S. (This is not a formal definition, these central sets will be linked to the notion of separators in Section 7.) In both examples, removing S from $\mathcal{G}_{CSC}$ creates three connected components C_1, C_2, C_3. It can be shown, using similar calculus than for the CSC with tree structure, that X_{C_i} given $X_{\bar{C}_i}$ depends only on a subset S_i of S. For instance, for the CSC (a), we have $p(x_2, x_3 \mid x_1, x_4, \ldots, x_7) = p(x_2, x_3 \mid x_4)$, and for the CSC (b) $p(x_4 \mid x_1, x_2, x_3, x_5, x_6) = p(x_4 \mid x_3)$. Each connected component can thus be eliminated in turn in $p(x)$, creating terms $a_{C_i \cup S_i}/a_{S_i}$. After elimination of the C_is, the last remaining term is a_S. Doing so we obtain the following decomposition of $p(x)$ for the two examples

$$\text{(a)} \quad p = \frac{a_{14} a_{234} a_{4567}}{a_4^2}$$
$$\text{(b)} \quad p = \frac{a_{123} a_{15} a_{26} a_{34}}{a_1 a_2 a_3}$$

In all the toy examples presented here, one can notice that the exponent β_I associated to a given specified marginal a_I ($I \in \mathcal{A}$) is always equal to 1 if $I \in \mathbb{G}(\mathcal{A})$ and negative or null otherwise. This property will be rigorously established in Section 7.

From these examples, one can get the intuition that if it is possible to establish an order to eliminate successively points or group of points of $\mathcal{V}$, by application of conditional independence induced by the specified marginals, we can derive the canonical exponents as functions of the CSC structure.

6. Elimination order and chordal graphs. The notion of elimination order has been defined formally in graph theory and has been linked with the notion of *chordal* graph ([6]). In this section, we present briefly these elements that will be essential to the establishment of an analytical solution of the ME problem.

Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ be a graph. A *clique* is a set of vertices such that any two vertices of this set are linked by an edge. If $c \subset \mathcal{V}$ is a clique, then any subset $c' \subset c$ is a clique

too. A *maximal clique* is a clique which is not strictly included in another clique. The set of maximal cliques in $\mathcal{G}$ is denoted $\mathcal{C}$.

A vertex is a *simplicial vertex* if its neighborhood (set of vertices linked to it by an edge) is a clique. This clique is then maximal. An ordering $\{i_1, \ldots, i_n\}$ of a graph with n vertices is an ordering of the vertices. For a given ordering, let us define $\mathcal{G}_\alpha$ as the subgraph induced by vertices $(i_k)_{k \geq \alpha}$. Then a *simplicial elimination order* is an ordering such that for any $1 \leq k \leq n$, i_k is simplicial in subgraph $\mathcal{G}_k$ ([18]). Graphs which present a simplicial elimination order has been characterised: there exists a simplicial elimination order if and only if the graph is *chordal*. A graph is said *chordal* if any loop of length > 3 is cut by an edge. This class of graphs is also referred to as decomposable or triangulated graphs ([16]).

As a preliminary result for propositions 7.1 and 7.2, we recall a property of graphical models ([19, 4]) with chordal graph $\mathcal{G}_{GM}$: the joint distribution can be expressed as a product of some marginal distributions.

LEMMA 6.1. *Let p by a graphical model on Ω indexed on $\mathcal{B}$. If the graph $\mathcal{G}_{GM}$ associated is chordal then*

$$p(x) = \frac{\prod_{c \in \mathcal{C}} p_c(x_c)}{\prod_{S \in \mathcal{S}} p_S(x_S)} \tag{6.1}$$

where $\mathcal{C}$ is the set of maximal cliques and $\mathcal{S}$ is a particular subset of the cliques of $\mathcal{G}_{GM}$.

Proof. We propose here only a sketch of the proof, to give to the reader the key elements. A rigorous demonstration can be found in ([19, 4]).

A nice property of chordal graph is that the maximal cliques can be decomposed into a join tree ([6]). A *join tree* (which is not unique in general) of a graph $\mathcal{G}$ is a tree, denoted $\mathcal{T}$ whose vertices are maximal cliques of $\mathcal{G}$. Edges have no specific definition, but must fulfill the *running intersection property*: if there exists a path between two vertices of $\mathcal{T}$, let say c and c', then points in $c \cap c'$ must be present in any maximal clique along this path. The notion is examplified on the two graphs of Fig. 5.2: a possible join tree for both graphs is drawn on Fig. 6.1.

An intersection of two maximal cliques of $\mathcal{G}$ linked by an edge in $\mathcal{T}$ is called a *separator*. It can be shown ([4]) that the set of all separators (repetition included) when running over all edges in $\mathcal{T}$ does not depend on the specific join tree built on $\mathcal{G}$.

In formula (6.1), the set $\mathcal{S}$ is the set of separators of the join tree of $\mathcal{G}_{GM}$. This formula can be recovered by applying the same iterative elimination procedure that in Section 5 to the junction tree instead of to $\mathcal{G}_{GM}$. Note that the same subset of $\mathcal{V}$ can be a separator several times in the join tree (see for instance node 4 on the join tree of Fig. 6.1 (a)). The set $\mathcal{S}$ is a list with possible repetitions. The number of occurrence of a separator S in $\mathcal{S}$ can be determined by building the junction tree. Moreover there is a relationship between these numbers and the Mœbius numbers of the decomposition of the entropy of a graphical model ([17]).□

7. Maximum entropy for chordal maximal CSC. Intuitively, if the graph $\mathcal{G}_{CSC}$ associated to a CSC is chordal, this will enable to built the maximum of entropy distribution recursively, by following the elimination sequence of the maximal cliques, leading to an expression of form (6.1). This expression depends on the marginal distribution on the maximal cliques. However, even if a generator of a CSC correspond to a clique in $\mathcal{G}_{CSC}$, it is not necessarily maximal. If $\mathbb{G}(\mathcal{A}) = \{\{1,2\},\{2,3\},\{3,1\}\}$, $\{1,2,3\}$ is a maximal clique of the graph $\mathcal{G}_{CSC}$ but does not correspond to any generator. Since in general a probability distribution on Ω_c cannot be uniquely defined

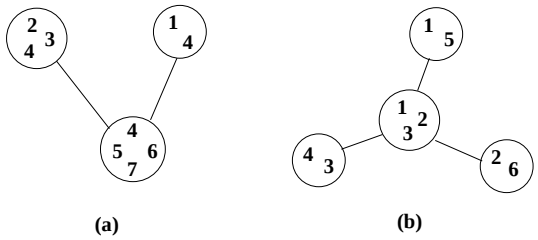

FIG. 6.1. *Examples of join tree: (a) join tree of graph on Fig. 5.2 (a), join tree of graph on Fig. 5.2 (b).*

from marginals on subsets of c, one can expect that if the generators of the CSC do not correspond exactly to the maximal cliques of $\mathcal{G}_{CSC}$, there will be difficulties to find the solution of the ME problem. We show now that the two properties, chordality and maximality of the generators, are sufficient conditions for an identification of the exponents in the expression of the ME under constraints on marginals.

Let us define the familly of *chordal maximal* CSC . Let $(\mathcal{A}, a_I)$ be a CSC , with $\mathbb{G}(\mathcal{A})$ the set of generators of $\mathcal{A}$. The CSC is *chordal maximal* if the graph $\mathcal{G}_{CSC}$ associated is chordal and if the maximal cliques of $\mathcal{G}_{CSC}$ correspond to the generators of $\mathcal{A}$.

THEOREM 7.1. *Let $(\mathcal{A}, a_I)$ be a chordal maximal CSC . If $\mathcal{C}$ and $\mathcal{S}$ are respectively the set of maximal cliques and the set of separators of the graph $\mathcal{G}_{CSC}$ associated to the CSC , then the distribution defined as*

$$p(x) = \frac{\prod_{c\in\mathcal{C}} a_c(x_c)}{\prod_{S\in\mathcal{S}} a_S(x_S)} \tag{7.1}$$

is the joint distribution with maximum entropy under constraints defined by $(\mathcal{A}, a_I)$.

Proof. Expression (7.1) is a decomposition matching (4.3): since separators are cliques of $\mathcal{G}_{GM}$, they are elements of $\mathcal{A}$, thus the a_S are known from the constraints set. By Lemma 6.1 this is the joint distribution of a graphical model with graph $\mathcal{G}_{GM}$ equal to $\mathcal{G}_{CSC}$ and marginal distributions on the maximal cliques of $\mathcal{G}_{GM}$ (the generators of the CSC) equal to the a_I. Thus p complies with constraints of the CSC . By Lemma 4.2, $\mathcal{H}(p)$ is a majorant of the entropy of any distribution satisfying to the constraints. As p belongs to this set, majorant is reached by p, and p is the distribution with maximal entropy under the contraints on marginals defined by $(\mathcal{A}, a_I)$. □

Conversely, any graphical model with chordality property can be defined as the solution of a ME problem with specification of a particular set of marginals.

THEOREM 7.2. *A graphical model with chordal graph $\mathcal{G}_{GM}$ is the distribution with maximum of entropy when marginals on the maximal cliques of $\mathcal{G}_{GM}$ are specified.*

Proof. Indeed, the joint distribution of the graphical model is given by expression (6.1). Let us consider the CSC $(\mathcal{A}, a_I)$ with generators equal to the maximal cliques of $\mathcal{G}_{GM}$. Since $\mathcal{G}_{GM} = \mathcal{G}_{CSC}$, this CSC is in the familly of chordal maximal CSC . By proposition 7.1, the graphical model is solution of the maximum of entropy under $(\mathcal{A}, a_I)$. □

8. Conclusion and discussion. Maximum entropy under constraints problem has a long history in statistical mechanics ([10, 11, 9]) and in information theory ([21, 3, 9]). The particular case where the available incomplete information is a set of marginals of an unknown joint distribution occurs in probabilistic reasonning ([23])

or statistical inference ([13]). Methods have been developped to solve numerically the optimisation problem (see for instance [13, 8, 23] and reference therein), even in the case of relaxed constraints ([7]). In this latter case, the authors prove that a relaxed version of maxent gives an almost best solution. However, under certain conditions on the set of constraints, it is possible to analytically express the maximum entropy solution as a function of the specified marginals. Literature related to this topic is somehow scattered in disciplines such as statistical mechanics, information theory, graph theory or inference in graphical models. In this article we gather and link results from these different fields. From this, we show that for a particular class of constraints set on marginal, the *chordal maximal coherent sets of constraints*, it is possible to derive analytically the distribution of the graphical model solution of the maximum entropy problem and to express the potential functions as functions of the constraints and the problem's structure.

It is not surprising that the notions of elimination order and chordal graph are key elements in this result. They have been long studied in graph theory for their application in bayesian networks inference ([14, 12]), or in constraint satisfaction problems ([5]). It is well known that when the moral graph of a bayesian network is chordal, there exist methods for exact inference, based on intelligent message passing algorithms ([14, 12]). When the graph of a Markov Random Field is a tree (particular case of a chordal graph), finding the configuration with Maximum a Posteriori probability is tractable using *maxsum* algorithms ([2][chapter 8]). Similarly, when the graph of the constraints in a constraint satisfaction problem is chordal, exact optimisation methods are available ([5]).

In this article, we have essentially explored the information theory part of the maximum entropy principle. As we mentionned earlier, this notion is also essential in statistical mechanics. Statistical mechanics are at the origin of a familly of methods, namely the Kikuchi or variational methods ([15, 24]), for the approximation of marginals of a complex joint distribution by ones simpler to compute. These methods are built as truncatures of the Mœbius decomposition of the complex distribution. Well known order 1 and order 2 elements of this familly are respectively the mean field and the Bethe approximations. They do not necessarily correspond to an approximation of the complet joint distribution which is normalised. One open question closely linked to the one addressed in this paper is: under which conditions is the variational approximation of a joint distribution exact? In that case, can it be derived by a maximum entropy principle? Intuitively, when a variational approximation correspond to a valid joint distribution, it is the joint distribution of a graphical model. If truncature is at order lower than the size of the largest clique, approximation can not be exact. But what if truncature is at order higher? If conditions for equality between a graphical model and its variational approximation can be established rigorously, then under chordality assumption, a variational approximation is the solution of a maximum of entropy problem under constraints on some marginals. We are currently investigating this issue. Such a result would open new directions to exploit variational methods not only as approximations of marginals (Generalised Belief Propagation algorithms for graphical models inference [24, 20]) but also of joint distributions.

9. Appendix. *Lemma 4.1. The joint probability distribution p with maximum of entropy compatible with a CSC $(\mathcal{A}, a_I)$ is a graphical model indexed by $\mathcal{A}$.*

Proof. The demonstration is straightforward when maximising the entropy under constraints using Lagrange multipliers. Let q be the solution of the ME problem. Each contraint has the form $q_I = a_I$, for $I \in \mathbb{G}(\mathcal{A})$ (contraints on elements which are

not generators are derived from these ones). The constraints are then (with $\bar{I} = \mathcal{V} \setminus I$)

$$\forall\, I \in \mathbb{G}(\mathcal{A}), \quad \sum_{x_{\bar{I}}} q\,(x_I, x_{\bar{I}}) = a_I(x_I)$$

If $\phi_I(x_I)$ is the Lagrange multiplier associated to generator I and state x_I, then the quantity to maximise is

$$H(q) - \sum_{I \in \mathbb{G}(\mathcal{A})} \sum_{x_i \in \Lambda} \phi_I(x_I)(\sum_{x_{\bar{I}} \in \Omega_{\bar{I}}} q\,(x_I, x_{\bar{I}}) - a_I(x_I))$$

Derivation leads to

$$\forall\, x \in \Omega, \quad \sum_{I \in \mathbb{G}(\mathcal{A})} \phi_I(x_I) = 1 + \operatorname{Log} q(x)$$

We obtain the following factored form for q:

$$\forall\, x \in \Omega, \quad q\,(x) = \lambda \prod_{I \in \mathbb{G}(\mathcal{A})} \psi_I(x_I)$$

where λ corresponds to the Lagrange multiplier associated to the constraint $\sum_x q\,(x) = 1$, and $\psi_I(x_I) = \exp \phi_I(x_I)$. We recover the expression of a graphical model indexed on $\mathbb{G}(\mathcal{A})$. □

Lemma 4.2. Let $(\mathcal{A}, a_I)$ be a CSC *, any distribution on Ω which can be expressed as follows:*

$$a(x) = \lambda \prod_{I \in \mathcal{A}} a_I(x_I)^{\beta_I}$$

with $\beta_I \in \mathbb{Z}$, and satisfies the CSC *is of entropy larger that any distribution q on Ω satisfying the* CSC *:*

$$\mathcal{H}(q) \leq \mathcal{H}(a)$$

Proof. Let us note $C(x) = q(x)/a(x)$. Then

$$\begin{aligned}\mathcal{H}(q) &= -\sum_{x \in \Omega} a(x)\, C(x) \operatorname{Log} a(x)\, C(x) \\ &= -\sum_{x \in \Omega} a(x)\, C(x) \operatorname{Log} a(x) - \sum_{x \in \Omega} a(x)\, C(x) \operatorname{Log} C(x)\end{aligned}$$

By convexity, $C(x) \operatorname{Log} C(x) \geq C(x) - 1$, so

$$\mathcal{H}(q) \leq -\sum_{x \in \Omega} a(x)\, C(x) \operatorname{Log} a(x) - \sum_{x \in \Omega} a(x)(C(x) - 1)$$

Since a and q are probability distributions, $\sum_{x \in \Omega} a(x)(C(x) - 1) = 0$, and

$$\mathcal{H}(q) \leq -\sum_{x \in \Omega} a(x)\, C(x) \operatorname{Log} a(x) \tag{9.1}$$

We establish now that

$$\sum_{x\in\Omega} a(x)\, C(x)\, \mathrm{Log}\, a(x) = \sum_{x\in\Omega} a(x)\, \mathrm{Log}\, a(x) = \mathcal{H}(a)$$

We use the following classical result of probability theory: if $p(x)$ is a probability distribution of a collection of variables indexed on $\mathcal{V}$, and f is a function only of the variables x_I for $I \in \mathcal{V}$ then,

$$\sum_{x\in\Omega} p(x) f(x_I) = \sum_{x_I\in\Omega_I} p_I(x_I) f(x_I)$$

Indeed, if we note $\bar{I} = \mathcal{V} \setminus I$

$$\begin{aligned}
\sum_{x\in\Omega} p(x) f(x_I) &= \sum_{x\in\Omega} p(x_{\bar{I}} \mid x_I) p(x_I) f(x_I) \\
&= \sum_{x_I\in\Omega_I} p(x_I) f(x_I) \sum_{x_{\bar{I}}\in\Omega_{\bar{I}}} p(x_{\bar{I}}) \\
&= \sum_{x_I\in\Omega_I} p(x_I) f(x_I)
\end{aligned}$$

since $\sum_{x_{\bar{I}}\in\Omega_{\bar{I}}} p(x_{\bar{I}}) = 1$. Using this result, we have the following equality

$$\forall\, I \in \mathcal{A}, \quad \sum_{x\in\Omega} a(x)\, C(x)\, \mathrm{Log}\, a_I(x) = \sum_{x\in\Omega} a(x)\, \mathrm{Log}\, a_I(x)$$

Since a_I is the marginal distribution of a and of q, both formula are equal to $\sum_{x\in\Omega} a_I(x_I)\, \mathrm{Log}\, a_I(x)$. Then

$$\begin{aligned}
\sum_{x\in\Omega} a(x)\, C(x)\, \mathrm{Log}\, a(x) &= \mathrm{Log}\, \lambda + \sum_{I\in\mathcal{A}} \beta_I \sum_{x\in\Omega} a(x)\, C(x)\, \mathrm{Log}\, a_I(x_I) \\
&= \mathrm{Log}\, \lambda + \sum_{I\in\mathcal{A}} \beta_I \sum_{x\in\Omega} a(x)\, \mathrm{Log}\, a_I(x_I) \\
&= \sum_{x\in\Omega} a(x)\, \mathrm{Log}\, \lambda \prod_{I\in\mathcal{A}} a_I^{\beta_I}(x_I) \\
&= \mathcal{H}(a)
\end{aligned} \tag{9.2}$$

Finally, from (9.1) and (9.2): $\mathcal{H}(q) \leq \mathcal{H}(a)$. □

References.

[1] C Berge. *Graphs and hypergraphs.* Elsevier, 1973.

[2] C. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer, 2006.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications. Wiley Interscience, New York, USA, 1991.

[4] R. Cowell, A. Dawid, S. Lauritzen, and D. Spiegelhalter. *Probabilistic networks and expert systems.* Statistics for Engineering and Information Science. Springer, 1999.

[5] R. Dechter. *Constraint processing.* Morgan Kaufmann, 2003.

[6] R. Diestel. *Graph theory.* Springer, 1997.

[7] M. Dudik, S. Phillips, and R. Schapire. Performance guarantees for regularized maximum entropy density estimation. In *Proceedings of the 17th Annual Conferenceon Computational Learning Theory*, 2004.
[8] S. A. Goldman. Efficient methods for calculating maximum entropy distributions. Technical report, master thesis, MIT EECS Department, 1987.
[9] A. Greven, G. Keller, and G. Warnecke, editors. *Entropy*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2003.
[10] E. T. Jaynes. Information theory and statistical mechanics, I. *Physical Review*, 106:620–630, 1957.
[11] E. T. Jaynes. Information theory and statistical mechanics, II. *Physical review*, 108:171–190, 1957.
[12] Finn V. Jensen. *Bayesian Networks and Decision Graphs*. Information Science and Statistics. Springer, 2001.
[13] J. Johnson, V. Chandrasekaran, and A. Willsky. Learning markov structure by maximum entropy relaxation. In *11th Inter. Conf. on AI and Stat*, San Juan, Puerto Rico, March 2007.
[14] M. Jordan. *Learning in graphical models*. The MIT Press, 1998.
[15] R. Kikuchi. A theory of cooperative phenomena. *Physical Review*, 81(6):988–1003, 1951.
[16] S. L. Lauritzen. *Graphical Models*. Clarendon Press, 1996.
[17] P. Pakzad and V. Anantharam. Estimation and marginalization using kikuchi approximation methods. *Neural Computation*, 17:1836–1873, 2005.
[18] C. Paul. *Parcours en largeur lexicographique : un algorithme de partitionnement. Application aux graphes et Generalisations*. PhD thesis, Université des Sciences et Techniques du Languedoc, 1998.
[19] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufman, 1988.
[20] A. Pelizzola. Cluster variation method in statistical physics and probabilistic graphical models. *Journal of Physics A: Mathematical and General*, 2005.
[21] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Ill., 1949.
[22] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, pages 1–305, 2008.
[23] J. Williamson. Maximising entropy efficiently. *Electronic Transactions in Artificial Intelligence*, 2002.
[24] J. Yedidia, W. Freeman, and Y. Weiss. Constructing free energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7):2282–2312, 2005.