An analysis of the impact of Kikuchi approximations

Schlüter Federico, Santana Roberto, Lozano José Antonio

Abstract

Factorizations of probabilistic distributions serve as condensed and efficient representations for modeling problems with uncertainty. In probabilistic graphical models, it is generally assumed that factorizations are consistent with a graphical structure which simplifies the probabilistic structure by exploiting the conditional independences present in the distribution. In this work we focus con Markov networks, a subtype of graphical models where the graphical structure is an undirected graph, and the factorization is characterized over the maximal cliques of the graph, according with the well-known Hammersley-Clifford Theorem. When using that factorization, factors may correspond to the maximal cliques and overlappings of a junction tree. The semantics of a junction tree guarantees that if marginals are consistent, whatever the accuracy of the approximation, the factorization will produce a valid probability distribution (e.g. the sum of the values associated by the factorization to all possible states of the system will add to 1). However, the number of junction trees, and thus the number of (consistent) factorizations that can be obtained from them is relatively small in comparison with all possible products of marginal distributions, including those not necessarily producing valid factorizations. In this work we investigate, from different perspectives, the quality of marginal factorizations as approximations of probability distributions. In particular, our analysis is focused on the Clique-based Kikuchi approximations (CBKA), a particular type of factorization in probability marginals. In this type of factorization, the marginals are completely determined by the independence graph. Our general goal is to identify whether, and under which conditions, the CBKA can produce accurate or acceptable approximations of an original probability distribution. We show that the class of CBKA includes a much larger set of approximations than those derived from junction trees (or equivalently, chordal graphs). Additionally, we show that the quality of CBKA can be measured by using the Kendall tau rank distance, a measure which considers scenarios where the values produced by the factorizations are mainly relevant to compare or rank the configurations of the system. We analyzed the quality of CBKA in terms of the Kendall distance and also in terms of the Kullback-Liebler divergence, and show that both measures are correlated. Our current results show that such correlation depends on the structure of the underlying distribution, and also on the strength of the dependences, e.g., α parameter of Dirichlet for synthetic distributions.

1 Introduction

Factorizations of probabilistic distributions serve as condensed and efficient representations for modeling problems with uncertainty. In simple terms, a factorization is a product of marginal factors that serves to approximate a Joint Probability Distribution (JPD). Usually, a factorization involves factors of much smaller size than that of the full JPD, guaranteeing that the factorizations are tractable in terms of the number of parameters needed to keep all marginal probabilities.

In probabilistic graphical models, it is generally assumed that factorizations are consistent with a graphical structure which simplifies the probabilistic structure by exploiting the conditional independences present in the distribution [Pearl, 1988]. The two branches of graphical models that are more commonly used are Bayesian networks and Markov networks [Koller and Friedman, 2009]. For the particular case of Bayesian networks, the structure is an acyclic directed graph, and the JPD factorizes over a product of conditional probability distributions in a compact and modular way. For Markov networks (MNs) the graphical structure is an undirected graph, and the factorization is characterized over the maximal cliques of the graph, according with the Hammersley-Clifford Theorem [Besag, 1974, Lauritzen, 1996]. When using that factorization, factors may correspond to the maximal cliques and overlappings of a junction tree. The semantics of a junction tree guarantees that if marginals are consistent, whatever the accuracy of the approximation, the factorization will produce a valid probability distribution (e.g. the sum of the values associated by the factorization to all possible states of the system will add to 1). However, the number of junction trees, and thus the number of (consistent) factorizations that can be obtained from them is relatively small in comparison with all possible products of marginal distributions, including those not necessarily producing valid factorizations.

In this work we investigate, from different perspectives, the quality of marginal factorizations as approximations of probability distributions. In particular, our analysis is focused on the Clique-based Kikuchi approximations (CBKA) [Santana et al., 2005], a particular type of factorization in probability marginals. In this type of factorization, the marginals are completely determined by the independence graph. Given an undirected graph, a unique CBKA is completely determined by computing the maximal cliques of the original graph, and then the cluster variation method is used to compute all possible overlappings. If the original graph is chordal, the obtained CBKA will be consistent with a junction-tree-based factorization. Our general goal is to identify whether, and under which conditions, the CBKA can produce accurate or acceptable approximations of an original probability distribution. We show that the class of CBKA includes a much larger set of approximations than those derived from junction trees (or equivalently, chordal graphs).

The following questions are addressed in our research[1]:

[1]The results obtained until now could be splitted in different of these lines. Actually, what

(i) We consider different measures of "quality" of the approximation. These measures include the *Kullback-Leibler* (KL) divergence, usually applied to measure distances between distributions, but also the *Kendall tau rank* distance, a measure used to evaluate the distances between rankings or permutations. The KL divergence is commonly used to measure the difference between two probability distributions. Thus, we compare the JPD from a CBKA with the original distribution in order to quantify how similar they are. Instead, the Kendall tau rank distance considers scenarios where the values produced by the factorizations are mainly relevant to compare or rank the configurations of the system. What is important in these cases is the capacity of the approximation to keep the relative order between the solutions of the space, not the distance to the original distribution in the KL sense. Notice that there exist multiple realistic situations in which we would like to be able to rank solutions with large number of attributes based on statistics computed from smaller subsets of these attributes. We analyze the quality of CBKA in terms of both quality measures, and show that they are correlated. Our current results show that such correlation depends on the structure of the underlying distribution, and also on the strength of the dependences, e.g., α parameter of Dirichlet for synthetic distributions.

(ii) We also investigate the relationship between general CBKAs and those that originate from chordal graphs using a measure of topological proximity among all possible approximations for a given number of variables n. The topological relationship is defined in terms of the subsumption relatioship among all possible undirected graphs for a given n. What we get from this topological relationship is a landscape of the CBKAs, where it is possible to identify which regions in the space of CBKA contain more chordal CBKAs and whether and how are they related. Although we do not develop in this work algorithms to learn CBKAs from the data, we hypothesize that this landscape could be instrumental for the design of this type of algorithms.

(iii) Finally, since CBKA are strongly dependent on the characteristic of the graph used to create the approximation, we are also interested in investigating the sensitivity of the CBKAs to structural errors. Specifically, we want to know how the chordality (or non-chordality) of the underlying structure affects to the robustness of the CBKA when using graphs which contain structural errors. For this, we consider two different types of errors that can be present in a structure: type-I errors and type-II errors. Type-I errors (a.k.a. false positives) correspond to additional incorrect edges added to the structure, assuming inexistent dependences between variables. Type-II errors (a.k.a. false negatives) correspond to incorrect independences assumptions, or spurious non-edges in the structure used

we are trying to do is to organize these results with respect to what could be the possible objectives of the work. We could discuss later what is the first thing we should try to finish

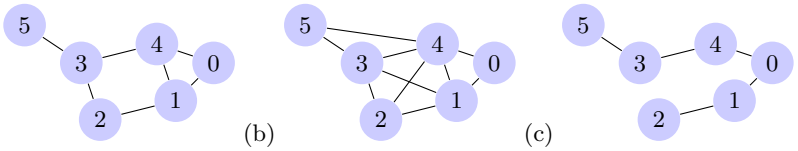

Figure 1: Description of the different types of errors. (a) Undirected graph representing the independence structure of a distribution, (b) Graph containing three type-I errors, (c) Graph that contains two type-II errors.

(i.e., assuming independence between two variables that are dependent in the underlying distribution). Figure 1 shows an example of the different types of errors. If the first graph is the underlying correct structure of a distribution, the second graph contains three type-I errors, and the third graph contains two type-II errors. Regarding the correctness of the distribution, type-I errors do not make incorrect assumptions over the functional form of the underlying distribution, because the independences in the structure still hold. That is, type-I errors may be mitigated by the numerical parameters of the graphical model. The problem with type-I errors is that they imply the use of an overcomplex model to represent the distribution. Instead, type-II errors add incorrect independence assumptions on the distribution, and these errors cannot be mitigated by the numerical parameters. Thus, type-II errors can invalidate statistical inference, leading to faulty conclusions.

This paper is organized as follows[2]: In the next section we present the notation and main concepts used for our analysis. Section 3 describes the quality measures of the approximation used, and analyzes the relationship between them. Section 4 defines the landscape of CBKAs and analyze it. In Section 5 we investigate the results for the analysis of the different types of errors. Finally, Section 7 lists the open questions and some ideas to extend this work.

2 Notation and background

Let $X = (X_0, \ldots, X_{n-1})$ denote a set of n discrete random variables. We will denote $x = (x_0, \ldots, x_{n-1})$ to an assignment of these variables. S will denote a set of indices in $\{0, \ldots, n-1\}$, X_S (respectively x_S) a subset of the variables of X (respectively x) determined by the indices in S, and $val(X_S)$ to the set of all possible values of X_S. We will work with positive distributions denoted by p. We use $p(X_S)$ to denote the marginal probability distribution of p on X_S. $p(x_S)$ will denote the marginal probability for $X_S = x_S$. We use $p(X_i|X_j)$ to denote the conditional probability distribution of X_i given X_j.

[2]This is a suggested organization for all the results during the visit, including those included here

An undirected graph $G = (V, E)$ is defined by a set of vertices V, and a set of edges E. An edge between nodes i and j will be represented by $i \sim j$. Given an undirected graph $G = (V, E)$, a clique in G is a fully connected subset of V. We reserve the letter C to refer to a clique. The collection of all cliques in G is denoted as $\mathcal{C}$. C is maximal when it is not contained in any other clique. C is the maximum clique of the graph if it is the clique in $\mathcal{C}$ with the highest number of vertices.

An undirected graph is said to be *chordal* if every cycle of length four or more has a chord. A fundamental property of chordal graphs is that their maximal cliques can be joined to form a tree, called the junction tree, such that any two cliques containing a node are either adjacent in the junction tree, or connected by a chain made entirely of cliques that contain that node.

2.1 Markov networks

Definition 1. *(Neighborhood): The neighborhood $N(X_i)$ of a node $X_i \in X$ is defined as $N(X_i) = \{X_j : X_j \sim X_i \in E\}$. The set of edges uniquely determines a neighborhood system on G.*

Definition 2. *(Boundary, bd): The boundary of a set of variables, $X_S \subseteq X$ is the set of variables $X \setminus X_S$ that neighbors to at least one variable in X_S. The boundary of X_S is denoted $bd(X_S)$.*

Definition 3. *(Closure, cl): The closure of a set of variables, $X_S \subseteq X$ is the set of variables $cl(X_S) = X_S \cup bd(X_S)$.*

Definition 4. *A Markov Network (MN) for X is a pair (G, Φ) where G is an undirected graph, and $\Phi = (\Phi_{C_1}, \ldots, \Phi_{C_c})$ is a set of nonnegative functions called neighbor potentials, one for each of the c maximal cliques in G. The distribution determined by the MN has the form:*

$$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} (\Phi_{x_C}) \tag{1}$$

and it is usually called a Gibbs Field with respect to the neighborhood system G. The normalizing constant Z is known as the partition function *and is given by $Z = \sum_X \prod_{C \in \mathcal{C}} (\Phi_{X_C})$.*

A MN fulfills a number of Markov properties. It satisfies that:

$$p(x_i, x/cl(x_i)|bd(x_i)) = p(x_i|bd(x_i)) \cdot p(x/cl(x_i)|bd(x_i))$$

This property is known as the *local Markov property* [Lauritzen, 1996].

2.2 Statistical inference with Markov networks

Belief propagation [Pearl, 1988, Aji and McEliece, 2000] has been traditionally used in statistical inference for obtaining a posteriori marginal probabilities in

graphical models. The approach is also useful to compute the most probable global states or system configurations [Nilsson, 1998]. The conection between belief propagation algorithms and free energy approximations in physical systems was presented in [Yedidia et al., 2003]. Yedidia et al. demonstrated that the fixed points of belief propagation algorithms coincide with the stationary points of Bethe's approximate free energy subject to certain consistency constraints. Bethe's approximation is known to be a special case of a more general class of approximations called Kikuchi free energy approximations. This result inspired an important amount of work [Minka, 2001a, Wainwright et al., 2001, Yuille, 2001, Tatikonda and Jordan, 2002, Heskes, 2003, Aji and Yildirim, 2003, Wainwright and Jordan, 2003, Pakzad and Anantharam, 2005, Yedidia et al., 2005, Wainwright and Jordan, 2008, Ravikumar et al., 2010, Weiss et al., 2012, Chen and Wang, 2012] that proposed new Generalized Belief Propagation algorithms (GBP), studied their conditions of convergence, and introduced new applications.

Belief propagation can be generalized by considering algorithms able to approximate marginals in graphs with cycles, also known as loopy graphs. The original belief propagation algorithm was theoretically proved to converge only in acyclic graphs. One common point in recent work on GBP is the reexamination and application of some old results achieved in statistical physics to the solution of inference problems in graphical models. Among these results is the Cluster Variation Method (CVM) [Kikuchi, 1951] introduced by Kikuchi in 1951 as a procedure to compute an approximation of the free energy. The algorithm, also known as the Kikuchi approximation of the free energy, has been used for the design of generalized propagation algorithms.

Generalized propagation algorithms assume that the structure of the graphical model is known, and organize the message passing steps for achieving an (eventually) good approximation of the desired marginal distributions. Methods from statistical physics are useful to determine the way message passing has to be organized and the conditions of convergence. The Kikuchi approximation has been used for two different although related problems: the problem of learning factorized approximations of probability distributions from data, and that of sampling from such type of approximations [Santana, 2005]. In this sort of application, the information about the structure of dependencies represented in the graphical model is totally absent or partial[3].

2.3 Kikuchi approximations

We define a region R of the independence graph $G = (V, E)$ to be a set $V' \subset V$. A graph region based decomposition is an asset of regions $\mathcal{R}$, and an associated set of *counting numbers* U which is formed by one counting number c_R for each $R \in \mathcal{R}$. c_R will always be an integer, but might be zero or negative for some R. In the Cluster Variation Method (CVM), $\mathcal{R}$ is formed by an initial set of regions $\mathcal{R}_0$ such that all the nodes are in at least one region of $\mathcal{R}_0$, and any

[3]In a new version of the manuscript this part has to be more related with the proposal we advance here in the paper (see To do section at the end of this manuscript)

other region in $\mathcal{R}$ is the intersection of one or more of the regions in $\mathcal{R}$. The set of regions $\mathcal{R}$ is closed under intersection, and can be ordered as a poset.

To be valid, a decomposition must satisfy a number of constraints relating the regions and the counting numbers. Inspired in the work by Yedidia et al. (2005) [Yedidia et al., 2005], we call this sub-problem as that of *finding a valid region based decomposition of a graph*. We say that a set of regions $\mathcal{R}$, and counting numbers c_R give a valid region based graph decomposition when for every variable X_i:

$$\sum_{\substack{R \in \mathcal{R} \\ X_i \subseteq X_R}} c_R = 1 \tag{2}$$

We will form the set $\mathcal{R}_0$ by taking one region for each maximal clique in G. As a result, all the regions $R \in \mathcal{R}$ will be cliques because they are the intersection of two or more cliques. We call this type of region based decomposition of undirected graphs a CBKA [Santana et al., 2005].

We define the Kikuchi approximation of the probability associated to a clique based graph decomposition, denoted as k as:

$$k(x) = \prod_{R \in \mathcal{R}} p(x_R)^{c_R}, \tag{3}$$

where $\mathcal{R}$ comes from a clique based graph decomposition. The overcounting numbers c_R are calculated using the following recursive formula:

$$c_R = 1 - \sum_{\substack{S \in \mathcal{R}_1 \\ S \supset R}} c_S \tag{4}$$

where c_S is the overcounting number of any region S in $\mathcal{R}_1$ such that S is a superset of R. c_R values corresponding to the initial maximal cliques are equal 1. If c_R is different from zero, the region is included in the clique based decomposition. [4]

2.4 Quality measures

As quality measures, we evaluate if $k(x)$ is a good approximation of the original JPD by two different measures: *(i)* The Kullback-Leibler divergence $KL(K|P)$; and *(ii)* the Kendall tau rank distance.

For two discrete probability distributions $p(x)$ and $q(x)$, the Kullback–Leibler divergence [MacKay, 2003] from $q(x)$ to $p(x)$ is defined to be

$$KL(P|Q) = \sum_i p(i) \log \frac{p(i)}{q(i)}. \tag{5}$$

In words, it is the expectation of the logarithmic difference between the probabilities $p(x)$ and $q(x)$, where the expectation is taken using the probabilities $p(x)$.

[4] to do: add definition of Normalized Kikuchi approximation $\tilde{k}(x)$

This is a measure of the difference between the two probability distributions $p(x)$ and $q(x)$. It is not symmetric in $p(x)$ and $q(x)$.

The other quality measure we are interested for evaluating the quality of the approximaated distributions obtained is the Kendall tau rank distance. This metric counts the number of disagreements between two ranking lists. The larger the distance, the more dissimilar the two lists are. The Kendall tau ranking distance between two lists is

$$K(\tau_1, \tau_2) = \sum_{\{i,j\} \in P} \bar{K}_{i,j}(\tau_1, \tau_2), \tag{6}$$

where P is the set of unordered pairs of distinct elements in τ_1 and τ_2, $\bar{K}_{i,j}(\tau_1, \tau_2)$ = 0 if "i" and "j" are in the same order in τ_1 and τ_2 $\bar{K}_{i,j}(\tau_1, \tau_2) = 1$ if "i" and "j" are in the opposite order in τ_1 and τ_2. For measuring the quality of the CBKA, we form the two lists τ_1 and τ_2 by sorting in descending order the configurations of the JPD, according with its respective probability, and then the number of disagreements is counted.

2.5 Storage cost of CBKAs

The storage cost of each CBKA clearly depends on the number of cliques, their size, and more important, the number of overlappings among the variables. The number of comparisons needed to find the first set of overlappings is $\frac{(\mu)(\mu-1)}{2}$, where μ is the number of maximal cliques. The process is repeated taking this maximum number of regions as a bound. A good estimator of the total number of regions can be $N_r = |C| \cdot \frac{(\mu)(\mu-1)}{2}$, where $|C|$ is the size of the maximum clique in the graph. However, for storing a CBKA we only need to save the maximal cliques of the graph, and then the rest of overlapping regions can be found recursively. Thus, we compute the cost of storing a CBKA as

$$cost(G) = \sum_{C \in \mathcal{C}(G)} \left(\prod_{val(X_C)} |X_C| \right), \tag{7}$$

where $configs(c)$ is the set of all the possible configurations of the variables of a clique and $|c|$ is the size of each clique.

In order to identify whether and under which conditions the CBKA can produce accurate or acceptable approximations of an original distribution, we consider the storage cost associated with a given quality. In Figure 2, a table of plots show the distribution of storage cost for binary problems with 4, 5 and 6 variables (in different columns). The green (red) bars correspond to the number of chordal (non-chordal) graphs that exist for each possible storage cost. As shown in the next section, the number of possible undirected graphs grows super-exponentially, as $2^{\binom{n}{2}}$, and the number of chordal graphs grows much more slowly than the number of non-chordal graphs. Thus, both chordal and non-chordal graphs have different distributions of costs, and their distributions change in a different way with the number of variables[5].

[5]Roberto asked me to include all our results. We need to complete this analysis over costs.

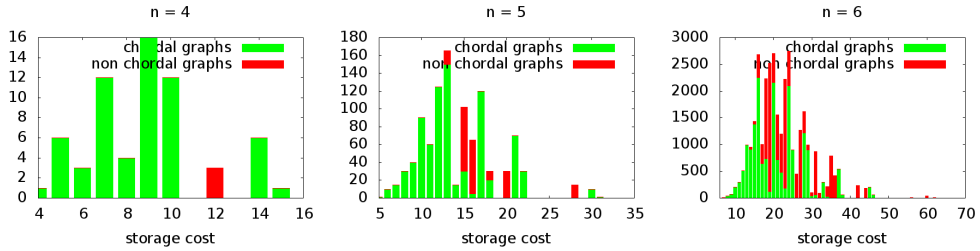

Figure 2: Cost of storing initial regions for n=4,5,6 (columns) for binary variables

3 Correlation between KL-divergence and Kendall distance measures

In this section we address the research question (i), about the relationship between two different measures of quality that can be used for CBKA. The specific question to investigate is about the correlation betweeen both quality measures. By looking at this correlation we expect to answer if it is possible to design novel methods based on the Kendall tau rank distance. Since in the literature the KL is probably as ubiquitous in algorithms for minimizing the energy, we want to know if it is possible to design methods based on choices of the energy designed from the definition of the Kendall distance.

The quality measures used are the Kullback-Leibler (KL) divergence and the Kendall tau rank distance. We show first how we generated the simulated distributions. Then, we describe a first experiment which shows how the correlation between both measures is clearly affected by the strength of the dependences when using arbitray specific independence structures. Finally, we present a systematic experiment for low-dimensional domains, where the Pearson's correlation coefficient between both quality measures is computed for all the possible cases.

3.1 Simulated distributions

For a specific domain size n, a JPD (joint probability distribution) is randomly generated. We consider distributions on n random binary variables. We assume that the independence structure of the distribution is given by a graph. For this, an arbitrary graph G with n nodes must be chosen. Then, the set with all the maximal cliques C is computed with the Bron and Kerbosh algorithm [Bron and Kerbosch, 1973].

I can optimize the code to generate the histograms for larger domains. I think we can show two graphs: one graph for the distribution over chordal graphs and other over non-chordal graphs. At each one, we can summarize the results with one curve for different n values (e.g., $n = \{4, 6, 8, 10, 12\}$). I included this in the TO DO list of the last section in this running paper

Once all the cliques have been found, the numerical parameters Φ for each corresponding clique factor is randomly drawn from a Dirichlet distribution. The Dirichlet distributions is used as a prior distribution for our discrete variables, in order to generate a multinomial distribution for each clique factor. The α parameter of the Dirichlet distribution is an input parameter that indicates the strength of the dependences between the factor variables. Then, at some point, the CBKA has to be computed for all the possible structures of the landscape.

3.2 Correlation for arbitrary cases

In this section we show a preliminar experiment performed in order to show how both measures, KL-divergence and Kendall tau rank distance, are correlated. The experiment consists in computing systematically the CBKA for all the possible graphs, in distributions where the independence structure is arbitrary selected. Finally, the KL-divergence and the Kendall tau rank distance is computed for each graph. Since the space contains non-chordal structures, we compute normalize the JPD before the KL computation.

Although we have results for many structures, in this report we only show three specific cases with $n = 6$ variables: the fully graph, a ring graph, and an empty graph. To produce weak, uniform and strong dependences for each structure, the original distribution were generated as indicated in the previous section by using $\alpha = \{0.1, 1.0, 100.0\}$ for the Dirichlet prior. Figure 3 shows three scatter plots for each class of dependence pattern investigated (one for each α value). The plots show chordal structures by a green circle, and non-chordal structures by a red triangle. In the x-axis, the structures are disposed by the KL-divergence obtained. In the y-axis, they are disposed by the Kendall tau rank distance.

When analyzing Figure 3, it is clear that for the three cases the two quality measures are correlated. For the structures with greatest KL-divergence, the Kendall tau rank distance seem to have its maximum value, and the same occur with the lowest KL-divergence structure. Additionally, the correlation seem to be sensitive to the structure, since the three cases exhibit a different shape of the correlation. Also, the correlation is clearly affected by the strength of the dependences in the simulated MNs, since the correlation is more shaped as well as α increases.

These first results gave us the insight that the correlation between both quality measures depend on the structure of the underlying distribution, and also on the strength of the dependences, i.e., α prior of Dirichlet distribution. In the next section, a more systematic experiment over all the possible structures for each domain size is shown.

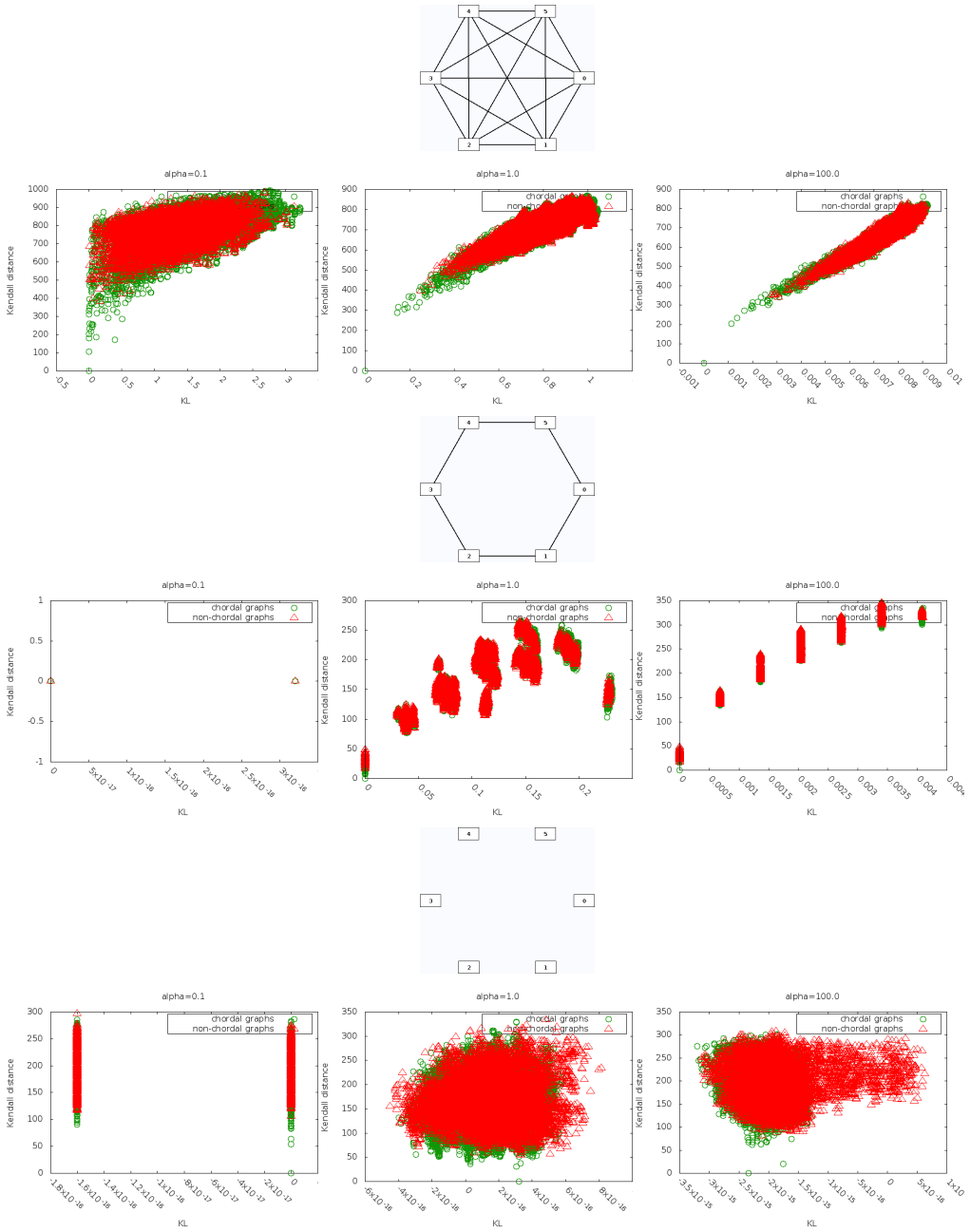

Figure 3: This scatter diagram graphs pairs the Kl-divergence (x-axis) with the Kendall distance (y-axis). Each column correspond to a different α value, i.e. stronger dependences in the underlying model. Each row show different study cases, that is different graph structures in the underlying distribution. The variables tend to be correlated for $\alpha \geq 1$, since the points will fall along a line. The higher the number of strong dependences in the underlying structure, the tighter the points hugs the line.

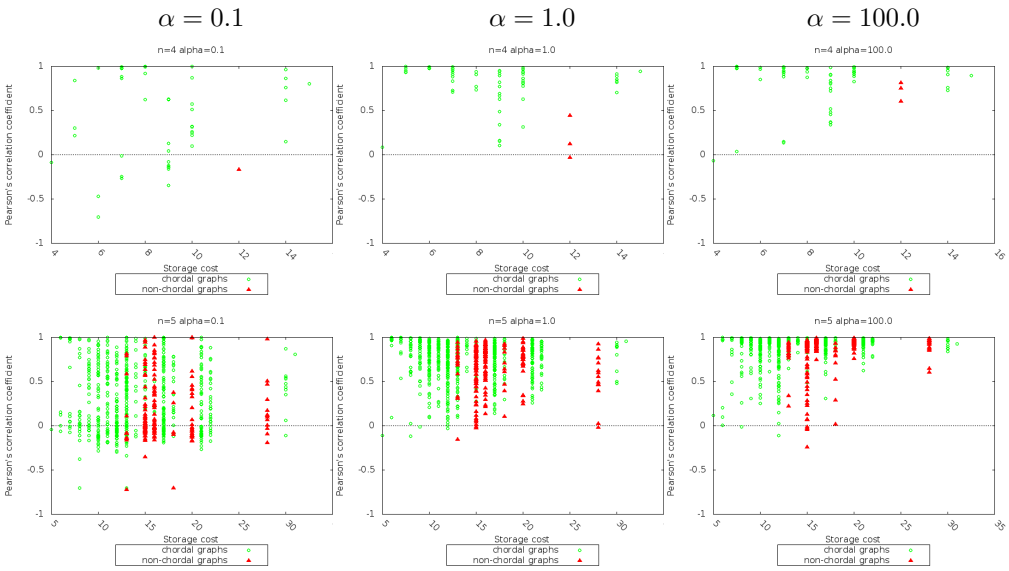

Figure 4: Pearson's correlation coefficient betweeen KL-Kendall distances for all the possible subgraphs ($n = 4$ in the first row, and $n = 5$ in the second row)

3.3 Pearsons correlation coefficient between KL-divergence and Kendall tau rank distance measures

In this section, the experiment made in the previous part is performed for all the possible arbitrary structures of a specific domain size. That is, a distribution has been created for each possible graph of size n, and then the KL-divergence and Kendall tau rank distance of all the possible CBKA have been computed and saved for each case. The goal is to analyze the correlation between both measures (KL-divergence and Kendal distance) for all possible graphs. For this, the Pearson product-moment correlation coefficient is used here as a measure of the linear dependence between both quality measures. Such coefficient takes values between $[-1, +1]$, where where 1 corresponds to total positive linear correlation, 0 is no linear correlation, and -1 is a total negative linear correlation.

Figure 4 shows the results of the experiment for $n = 4$ (first row), and for $n = 5$ (second row). For each domain size the experiment has been performed by using $\alpha = \{0.1, 1.0, 100.0\}$ for the Dirichlet prior. The plots dispose in the x-axis each possible distribution by the storage cost of the underlying structure, and show the Pearson's correlation between both quality measures in the y-axis. Here, the Pearson's coefficient is shown with a green circle for distributions where the underlying structure is chordal, and a red triangle for distributions where the underlying structure is non-chordal.

Clearly, the two different quality measures are correlated, since in all the plots the coefficient is positive more often than not. For $n = 4$, only three

distributions with non-chordal structures exist, and for this case the correlation is near to zero when $\alpha = 0.1$, and positive for greater α values. For $n = 5$, 20% of the CBKAs have a non-chordal structure. There are some isolated cases where there is no linear correlation (near to zero), but the value is positive or negative for most cases. In general, there is a clear tendency to appear more positive correlations as well as α increases[6].

It confirms our hypothesis that the correlation is sensitive to the structure. Also, the correlation is clearly affected by the strength of the dependences in the simulated MNs (α value)[7]. This results indicates that the tau rank distance obtained from CBKAs can be used as a surrogate of the KL divergence measure, when medium and strong dependencies exist in the underlying model.

4 A landscape of factorizations

In this section we address the research question (ii) of the introduction, concerned about the relationship between general CBKAs and those that originate from chordal graphs using a measure of topological proximity among all possible approximations for a given number of variables n. We define the topological relationship in terms of the subsumption relatioship among all possible undirected graphs for a given n. In this way, we get a landscape of the CBKAs, where it is possible to identify which regions contain more chordal CBKAs and whether and how are they related. In this work we do not develop algorithms to learn CBKAs from the data, but we hypothesize that this landscape could be instrumental for the design of this type of algorithms.

Let $\mathcal{G}$ denote the space of all the possible undirected subgraphs with n nodes. The size of the space $\mathcal{G}$ grows super-exponentially, as $2^{\binom{n}{2}}$. Instead, the number of chordal graphs grows much more slowly [Wormald, 1985], following the formula $a_n = c_n + \frac{1}{n}\sum_{k=1}^{n-1} k\binom{n}{k} c_k a_{n-k}$. Table 1 shows the percentage of chordal graphs over the total size of the graphs space, and it can be seen that its proportion tends to zero for relatively small problems ($n \geq 8$). Figure 6 illustrates such proportion in comparison with the percentage of non-chordal graphs (which oppositely, tends to one). As an additional example, the complete landscape for domains of size $n = \{4, 5, 6\}$ are shown in the hypercubes of Figure 5, where each node is a graph of the landscape, blue nodes are the chordal graphs, red nodes are non-chordal graphs, the structure on the top of the figure is the fully graph, the structure on the bottom is the empty graph, and edges connect neighbor graphs (graphs that only differ in one edge). Such hypercubes can be seen as alternative, more formal view of the fitness landscape of CBKAs, particular to search algorithms [Jones, 1995].

[6] It would be interesting to study why in some particulra cases the correlation appear to be negative.

[7] This experiment is really expensive to perform for higher domains, but I can optimize the code to generate similar results for higher domains

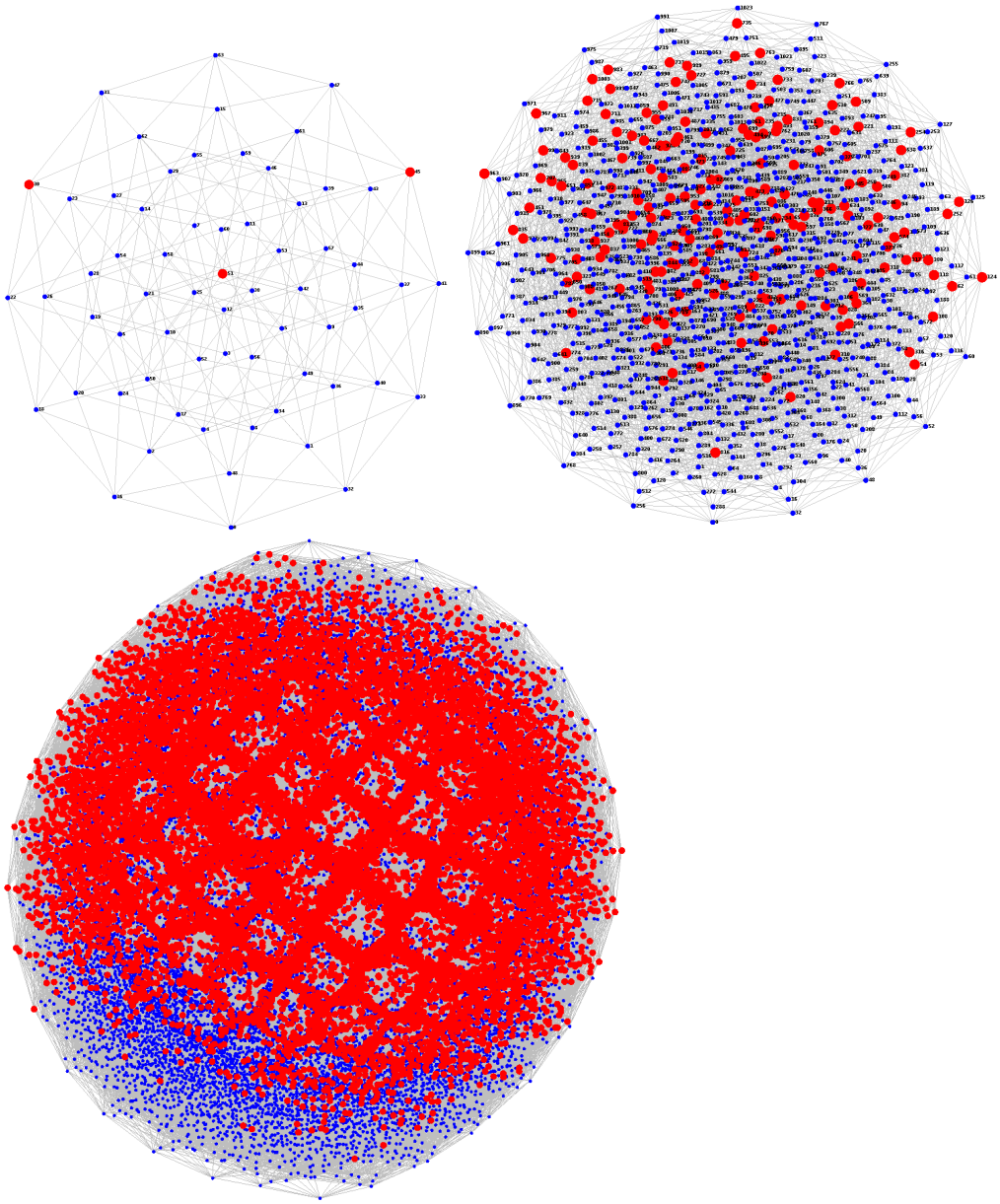

Figure 5: Distribution of non-chordal graphs (big, red) and chordal graphs (small, blue) on the landscape of graphs with n=4,5,6 variables

n	$2^{\binom{n}{2}}$	chordal graphs	percentage (%)
1	1	1	**100.000**
2	2	2	**100.000**
3	8	8	**100.000**
4	64	61	**95.310**
5	1024	822	**80.270**
6	32768	18154	**55.400**
7	2097152	617675	**29.450**
8	268435456	30888596	**11.500**
9	68719476736	2192816760	**3.190**
10	35184372088832	215488096587	**0.610**
11	3,6028797018964E+016	28791414081916	**0.070**
12	7,37869762948382E+019	5165908492061926	**0.007**

Table 1: Number of possible CBKAs, chordal graphs and percentage of chordal graphs over the total for $n \in \{1, \ldots, 12\}$.

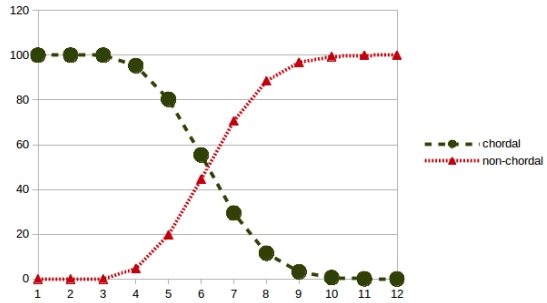

Figure 6: Percentage of chordal and non-chordal graphs in the complete space of $2^{\binom{n}{2}}$ graphs for graph sizes $n \in \{1, \ldots, 12\}$.

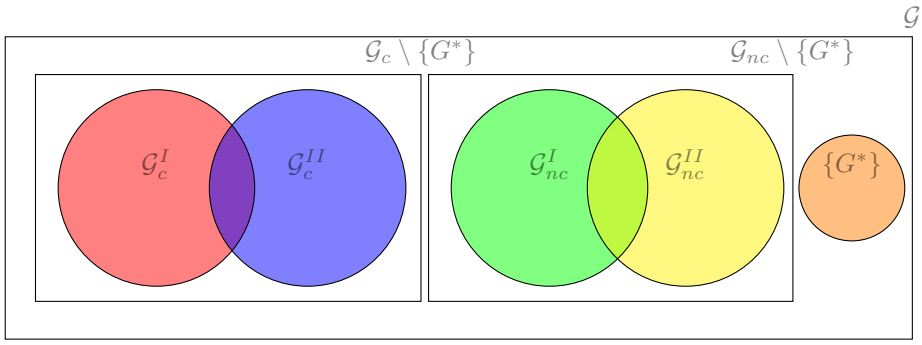

Figure 7: Space of all posible graphs $\mathcal{G}$, partitioned in chordal and non-chordal classes, and grouped by type I and type II errors, together with the solution.

We will denote $\mathcal{G}_c$ to the set of all chordal graphs, $\mathcal{G}_{nc}$ to the set of all non-chordal graphs, and G^* to the structure of the underlying distribution. Then, we know that $\mathcal{G} = \mathcal{G}_c \cup \mathcal{G}_{nc}$, and $\mathcal{G}_c \bigcap \mathcal{G}_{nc} = \emptyset$. Moreover, when choosing a specific structure $G \in \mathcal{G}$ to compute the CBKA, it can have type-I errors and/or type-II errors. Thus, we will call $\mathcal{G}_c^I$ to the chordal graphs with type-I erros, $\mathcal{G}_c^{II}$ to the chordal graphs with type-II erros, $\mathcal{G}_{nc}^I$ to the non-chordal graphs with type-I erros, and $\mathcal{G}_{nc}^{II}$ to the non-chordal graphs with type-II errors.

The landscape of all CBKAs can be partitioned as follows:

$$\mathcal{G} = \{G^*\} \cup \mathcal{G}_c^I \cup \mathcal{G}_c^{II} \cup \mathcal{G}_{nc}^I \cup \mathcal{G}_{nc}^{II}. \tag{8}$$

The Venn diagram of Figure 7 illustrates this partition. Note that the size of the circles in such diagram is not proportional to size of the sets. In fact, the respective size of such sets depends on n and the specific solution structure G^*.

5 Sensitivity of the clique-based Kikuchi Approximation

In this section we address the research question (iii) of the introduction, concerned about the sensitivity of the CBKAs in terms of their chordality, and its robustness to type-I errors and type-II errors. For this, we present in this section an experiment for answering the following questions:

(a) Should we expect a different quality for CBKAs when $G^* \in \mathcal{G}_c$, in contrast for those cases when $G^* \in \mathcal{G}_{nc}$?

(b) When G^* is unknown, is it always better to choose a chordal graph than a non-chordal graph? Or are there some cases where non-chordal graphs exhibit better performance?

(c) Is there some measure correlated to the quality of CBKA? (for example, some aspect of chordality, number of edges, number of cycles, etc.)

(d) When computing CBKAs with a structure G, can we say that type-I errors have a different impact than type-II errors in the quality of the approximation? Could we see some proportion? (How many type-I errors are equivalent in terms of KL or Kendall distance to type-II errors?)

5.1 Approximations from a MN structure

This section presents an experiment for measuring the quality of chordal an non chordal structures when the underlying distribution have a specific MN structure. For this, for an arbitrary graph, a MN were generated as explained in Section 3.1. Then, the CBKA has been computed for all the possible subgraphs of the landscape $G \in \mathcal{G}$. For each subgraph, the normalized KL-divergence and the Kendall tau rank distance measures were computed from the original distribution. Since the number of subgraphs grow exponentially, only the CBKA that are found as Pareto optimal solutions are shown. The Pareto optimal solutions were selected in terms of the KL over the storage cost, and in terms of the Kendall tau rank distance over the storage cost. By restricting attention to the set of choices that are Pareto-efficient, we can analyze the tradeoffs within this set, rather than considering the complete landscape of CBKA.

Figure 8 shows the results for an experiment for domains with $n = 6$ variables. We generated a Markov network for the 4 different graphs shown in the first column of each row. The CBKA that are found as Pareto optimal solutions in terms of the KL over the storage cost are shown in the second column. The third column show the Pareto optimal solutions in terms of the Kendall tau rank distance. These plots show only the Pareto optimal graphs, sorted by their storage cost in the x-axis. Those results can answer question (a) by looking if the Pareto optimal graphs are chordal (green circles) or non-chordal (red triangles). The graph of the third column shows Pareto optimal CBKAs, but in terms of the Kendall tau rank distance to the original distribution.

For the graphs of the first and fourth row of Figure 8, the underlying structure is chordal ($G^* \in \mathcal{G}_c$), and the graphs of the second and third row, the underlying structure is non-chordal ($G^* \in \mathcal{G}_{nc}$). When analyzing these results, it can be seen that the best solution is in most cases a CBKA over a chordal graph (green dots has lower KL), but also appear several interesting cases, where the best solution is non-chordal (red triangles). In the fourth row, all the structures have $KL = 0$. This was an expected result, that demonstrates that the type-II errors have an impact in the CBKA approximation, but type-I errors do not (at least using the marginals of the distribution, instead of approximate beliefs). The results in the second column correspond to Pareto optimal solutions in terms of Kendall tau rank distance. According to the correlation shown in the previous section, a similar behaviour can be seen for almost all the cases.

For $n = 6$, the percentage of chordal and non-chordal graphs is 56% and 44%, respectively (see Table 1). For larger domains, the proportion of chordal graphs is smaller. In Figure 9 the same experiment is shown for some graphs with $n = 8$. In this case, since it is unfeasible to compute the CBKA for all the

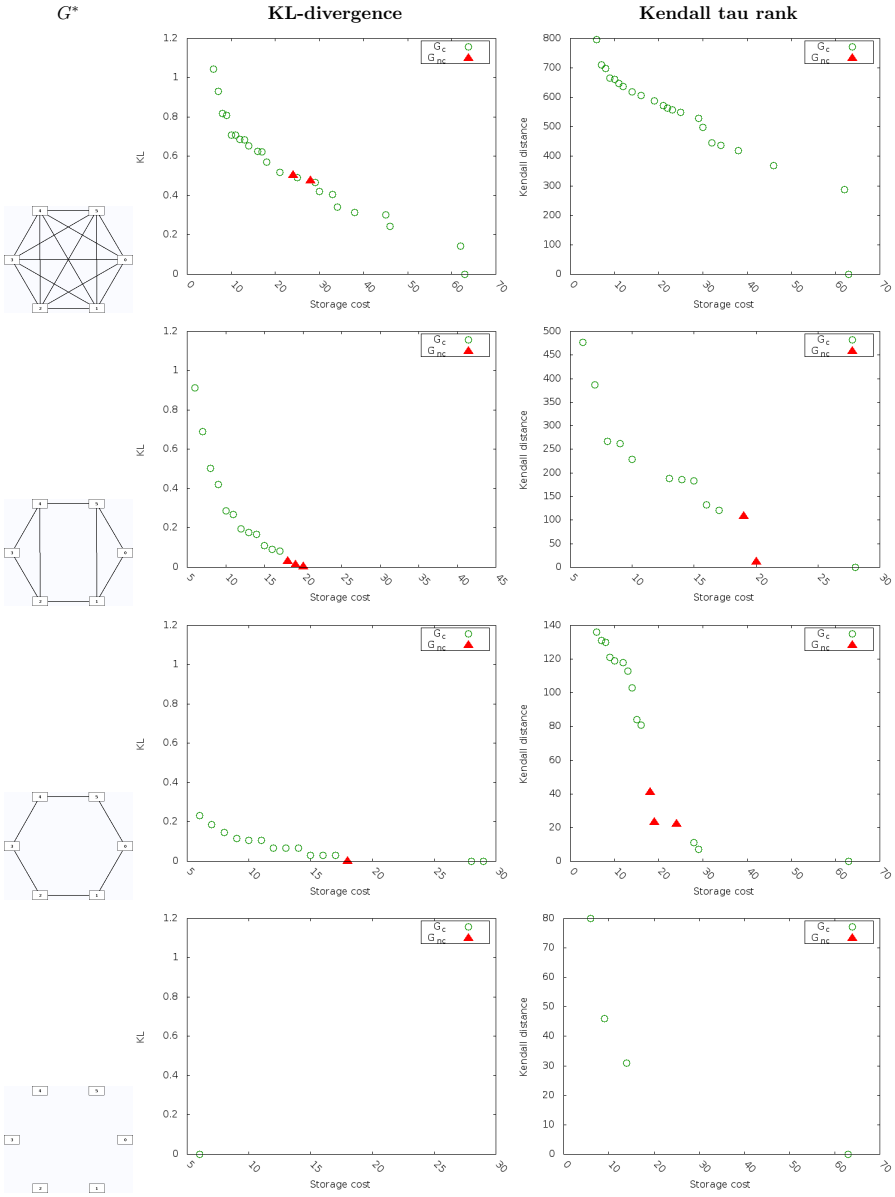

Figure 8: Pareto optimal solutions over the complete landscape of CBKA for different MNs with $n = 6$. Green circles are chordal Pareto optimal solutions, and red triangles are non-chordal Pareto optimal solutions. The second column shows the Pareto optimal solutions of the KL divergence over the storage cost. The third column shows the Pareto optimal solutions of the Kendall tau rank distance over the storage cost.

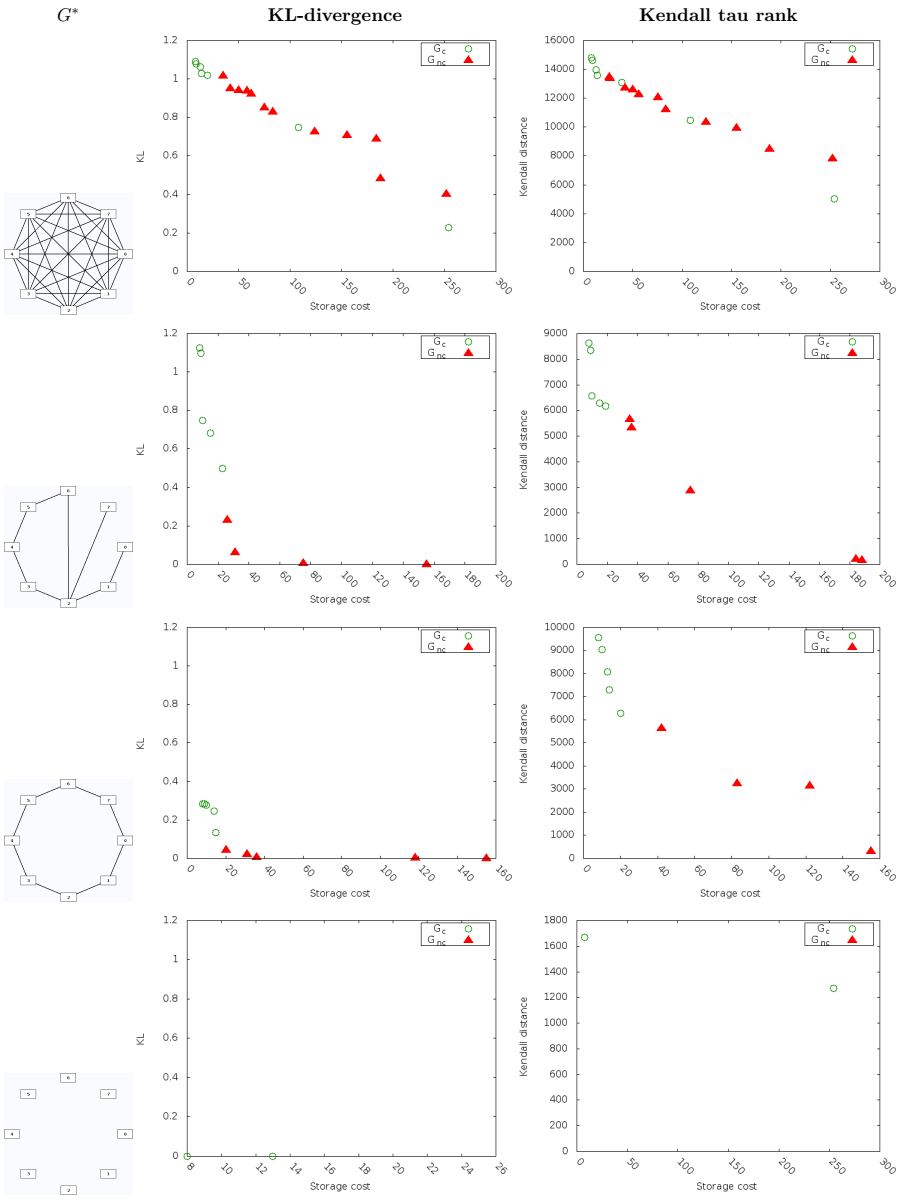

Figure 9: Pareto optimal solutions over a sample of the landscape of CBKA for different MNs with $n = 8$. Green circles are chordal Pareto optimal solutions, and red triangles are non-chordal Pareto optimal solutions. The second column shows the Pareto optimal solutions of the KL divergence over the storage cost. The third column shows the Pareto optimal solutions of the Kendall tau rank distance over the storage cost.

possible graphs, we uniformly sampled structures from $\mathcal{G}^8$. Interestingly, the trends are similar to the case for $n = 6$, but here the proportion of chordal and non-chordal graphs is 12% and 88%, respectively. In this case, more non-chordal cases appear in the list of Pareto optimal structures.

Regarding questions (a) and (b), our results does not show a clear difference between cases when $G^* \in \mathcal{G}_c$, in contrast with the cases where $G^* \in \mathcal{G}_{nc}$. For chordal graphs, when solution is the fully structure it is allways convenient to select chordal graphs, but it is not the same case for the empty structure. For non-chordal graphs, in some cases it seems to be convenient to select non-chordal structures, but we need a major insight about what are these specific cases. It seems that the impact of the CBKA is more related to the complexity of G^* (storage cost), and the number of type-II errors that the candidate structures can have.

Regarding question (c), it is possible that the number of cycles affect the quality of the CBKA. The results for the fully structure (that has a lot of cycles) show that the CBKA could be a worse approximation, since there is a clear trend to decrease the KL as well as the storage cost increases. The graphs in the third row shows that as well as the type-II errors increases the KL increases, which seem to be related with the cycles that are broken when type-II errors are introduced. Then the CBKA tends to be worse. For the empty structure, since there are no cycles the KL is always zero. For the non-chordal structures of Figures 9, it can be seen also that the cycles affect the CBKA. For the graph in the second row, we have also a worse approximation than that for the graph of the third row, because the KLs are greater.

Until now, the results do not cover all the specific cases, and a more systematic experiment is required to better answer our questions. But there are some clear trends. Additionally, in our experiments, the minimal triangulation graph is already computed and stored for non-chordal graphs in the Pareto set. They are not shown here, but it is of our interest in the future to compare the quality of the non-chordal CBKA with its corresponding minimal triangulated chordal graph, in order to check if there are improvements in quality and/or storage cost. The algorithm used for triangulation is extracted from [Berry et al., 2010], where the authors claim that although a minimum triangulation is NP complete, computing a minimal triangulation can be done in $O(nm)$ time (n is the number of vertices and m is the number of edges).

6 Related work

Below is a list of papers related to this research. They are sorted chronologically, with a brief description of the major contributions, and how each of them is related to the research questions of this work[9].

[8] In this experiment, we simply generated 10 structures for each possible cardinality, or number of edges

[9] Until now, I simple scanned these papers. I need more time to understand each of these papers more deeply, and to classify them correctly according to our current work

A work of interest is [Pelizzola, 2005], a review on the problem of the minimization of the variational free energy, which arises in the cluster variation method. This work describes the cases where the CVM is known to be exact. The first case is due to the topology of the underlying graph, i.e., when the solution graph is tree-like. Also, they describe the issue of *realizability* (the possibility of reconstructing a global probability distribution from the marginals predicted by the clustering variation methods) and consider cases where the form of the Hamiltonian makes an exact solution feasible with the CVM.

In [Pakzad and Anantharam, 2005], the problem treated is that of finding the marginals of a product distribution[10]. Our problem is different, given the marginals of a joint distribution, we want to know how accurate is a given product distribution, the one determined by the Kikuchi approximation from the original marginal distributions given (the original regions that correspond to maximal cliques of the graph). The questions addressed are two: 1) Finding one or more marginals. 2) Finding the partition function for the distribution. In this work we consider a related but different problem, given a distribution, not necessarily decomposable, what combination of its marginals produces a good approximation of the function given some predefined criterion. Loopy belief propagations and other message passing algorithms are used to find marginal distributions or the most probable configuration of a state. In the general case, it is not possible to estimate the quality of the approximation they produce. The rationale of applying message-passing algorithms, as done by Yedidia, is to obtain better approximations to the marginals and the partition function. The connection between the Statistical Physics and Kikuchi approximation given by the Pakzad adn Anantharam is based on a particular choice of the energy $E_r(x_r) = -log(\alpha_r(x_r))$ that leads to $F(b) = KL(b||B) - log(Z)$. One open questions is whether is possible to find another choice of the energy leading to a particular expression for the Kendal metric or other measures of quality of the distribution. In terms of the authors notation, the collection of values associated to the regions $\{b_r(x_r), r \in R\}$ are probability distributions. When they are marginals of another probability distribution they are called R-marginals. We work with R-marginals.

Another work of interest is [Heskes, 2006], an approach proposed to explicitly minimize the Kikuchi and Bethe free energy. This work tackles a specific problem of loopy and generalized belief propagation. Since these inference algorithms do not always converge to a stable fixed point, finding such a minimum then becomes a possibly non-convex constrained minimization problem. Thus, this work proposes an approach to solve this non-convex problem through sequential constrained minimization of convex bounds on the Kikuchi free energy. For this, the sufficient conditions for the Kikuchi free energy to be convex are discussed.

Another work of interest for this research question is [Komodakis et al., 2011]. This work introduces a framework for discrete MRF-based optimization in the computer vision problem. For MRFs, there are two classes of methods: those

[10]Roberto, he copiado y pegado las notas que tomaste sobre este paper y las he puesto aquí.

based on graph-cuts, and those based on message-passing. The latter class had a significant advance with the introduction of the tree-reweighted message passing algorithms [Kolmogorov, 2006, Wainwright et al., 2005]. The framework is based in the use of *dual decomposition*, one of the most powerful and widely used techniques in optimization. A projected subgradient scheme is used for computing a solution to a difficult or large optimization problem by first decomposing it into a set of easier subproblems and then combining the subproblems solutions. This leads to a message passing algorithm that generalizes tree-reweighted message passing methods and has stronger theoretical properties.

In [Welling et al., 2012] the authors stablished some connections between Generalized Belief Propagation algorithms [Yedidia et al., 2005] and EP algorithms for approximate Bayesian inference [Minka, 2001b]. The problem that they are interested to solve is how to choose an appropriate approximation structure. In this work, a framework called 'Structured Region Graph' is proposed for producing high-quality approximations with a user-adjustable level of complexity. The formalism proposed allows to choose good approximation structures to do inference with generalized belief propagation.

In [Korč et al., 2012], the problem of inference in a graphical model with binary variables is considered. The authors propose a method for computing marginal probabilities and also to use MAP inference, called the Discrete Marginals technique. In their method, approximate marginals are obtained by minimizing an objective function with unary and pairwise terms over a discretized domain. They propose two ways to set up the objective function: by discretizing the Bethe free energy and by learning it from training data. In their results, for certain types of graphs a learned function can outperform the Bethe approximation.

[Loh and Wibisono, 2014] is one more recent work, also focused in the concavity of reweighted Kikuchi approximation. This work proposes an objective function which is a reweighted version of the CBKA for estimating the log partition function of a distribution defined over a region graph. Sufficient conditions for the concavity of the function are stablished. When the region graph has only two layers (i.e., Bethe approximation), the sufficient conditions for concavity are also necessary. Also an explicit characterization of the polytope of concavity is provided, in terms of the cycle structure of the region graph. The authors claim that future research must include a better understanding of the approximation guarantees.

A recent related work is [Weller, 2015]. This work also considers the problem of inference in undirected graphical models, for the binary pairwise case. In this work the authors demonstrate that several recent results on the Bethe approximation may be generalized to a broad family of related pairwise free energy approximations with arbitrary counting numbers. The approximation error is analized, in order to explain the empirical success of the Bethe approximation.

Another recent work of interest for this research is [Lin et al., 2016]. In this paper the problem is how to do efficient inference in Bayesian networks with large numbers of densely connected variables. The paper presents an algorithm called Triplet Region Construction (TRC) for approximate inference which is

composed by three sub algorithms: *(i) Outer Region Identification (ORI)*: it allows to identify outer (redundant) regions in a first step to construct a valid region graph. The algorithm identifies largest regions by considering conditional independences derived from local correlations. The algorithm search for regions that satisfy 2 constraints. The first is the *perfect correlation property* which says that the sum of all regions overcounting numbers is 1. The second is the max-ent normal property, which says that the constrained region-based entropy achieves its maximum when all the beliefs are uniform. *(ii) Region Graph Binary Factorization (RGBF)*: This algorithms decomposes the region graph into an equivalent and numerically stable alternative. This is because for high tree-width factorized models the region-based algorithms suffer from numerically instability problems when performing inference. *(iii) Concave Convex procedure (CCP)*: This is a known alternative to Generalized Belief Propagation (GBP) to minimeze the Kikuchi free energy. In contrast with GBP, CCP guarantees convergence but it is numerically unstable for large models. The authors claim that the resulting TRC algorithm is guaranteed to converge and that it reduces the clustering complexity for factorized models from worst case exponential to polynomial.

7 Open questions and future work

Next, a list of tasks to be done in order to extend the experiments and analysis of the previous experiments.

- In a new version of the manuscript, Section 2.2 has to be more related with the proposal we advance.
- Regarding the costs histograms of Figure 2, we need to complete the analysis. It would be useful to find the function that fit these curves, in order to generate it automatically for higher domains, and perhaps for using it in learning algorithms. I can optimize the code to generate the histograms for larger domains. I think we can show two graphs: one graph for the distribution over chordal graphs and other over non-chordal graphs. At each one, we can summarize the results with one curve for different n values (e.g., $n = \{4, 6, 8, 10, 12\}$). We should investigate if there are known ways to compute how many graphs among the $2^{\binom{n}{2}}$ contain a particular clique, and from there derive the cost of all possible subgraphs.
- Regarding the Pearson's correlation scatter plots of Figure 4, it is required to optimize the implementation to generate the plots for larger domain sizes.
- The experiments of Section 5 have several issues. We need a more systematic methodology, since we cannot justify our arbitrary choices of the structure corresponding to the underlying distribution. Besides, we need an alternative methodology to extend the conclusions for larger domain

sizes. We can select also some benchmark networks of interest for the community as arbitrary structures.

- In the experiments of Section 5, the minimal triangulation graph is already computed and stored for the non-chordal graphs in the Pareto set. They are not shown here, but it is of our interest in the future to compare the quality of the non-chordal CBKA with its corresponding minimal traiangulated chordal graph, in order to check if there are improvements in quality and/or storage cost.

Finally, the following open questions arise from this work[11]:

- What are the properties of CBKAs learned from marginal distributions? Our results can answer partially this question. The fact that the Kendall tau rank distance is correlated to KL divergence is a specific property of the CBKA method? Our results in Section 5 showing several non-chordal structures in the Pareto optimal set demonstrate that the class of CBKA includes a much larger set of approximations than those derived from junction trees (or equivalently, chordal graphs)... what property of CBKA allows us to do this affirmation?

- Under which conditions a CBKA is a probability distribution? We used normalized KL in our experiments, but it is not tractable for larger domains. How can we tackle this problem?

- Regarding the correlation between KL divergence and Kendall tau rank distance, it is possible to take advantage of this for designing novel methods based on the Kendall tau rank distance. In the literature, the KL is probably as ubiquitous in algorithms for minimizing the energy. Can we design methods based on choices of the energy designed from the definition of the Kendall distance ?

- Regarding the landscape of CBKA proposed in Section 4, how can we use it for designing algorithms for learning CBKA? What criteria could we use to determine which structures are good?

- What problems should we focus? (options are: classification tasks, MAP inference tasks, marginal computations tasks, partition function computation, etc.)

References

[Aji and McEliece, 2000] Aji, S. M. and McEliece, R. J. (2000). The generalized distributive law. *IEEE Transactions on Information Theory*, 46(2):325–343.

[11] Jose Antonio, Roberto, creo que seria interesante que tomen nota de las posibles preguntas abiertas que tengan ustedes al leer esto, para agregarlo en esta lista y sumarlo a nuestra discusion

[Aji and Yildirim, 2003] Aji, S. M. and Yildirim, M. (2003). *Mathematical Systems Theory in Biology, Communications, Computation and Finance Series: The IMA Volumes in Mathematics and its Applications, Vol. 134*, chapter Belief propagation on partially ordered sets, pages 275–300. Springer.

[Berry et al., 2010] Berry, A., Pogorelcnik, R., and Simonet, G. (2010). An introduction to clique minimal separator decomposition. *Algorithms*, 3(2):197–215.

[Besag, 1974] Besag, J. (1974). Spacial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*.

[Bron and Kerbosch, 1973] Bron, C. and Kerbosch, J. (1973). Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577.

[Chen and Wang, 2012] Chen, S. and Wang, Z. (2012). Acceleration strategies in generalized belief propagation. *IEEE Transactions on Industrial Informatics*, 8(1):41–48.

[Heskes, 2003] Heskes, T. (2003). Stable fixed points of loopy belief propagation are minima of the Bethe free energy. In *Advances in Neural Information Processing Systems, (NIPS-2002)*, volume 14, pages 343–350. MIT Press.

[Heskes, 2006] Heskes, T. (2006). Convexity arguments for efficient minimization of the bethe and kikuchi free energies. *J. Artif. Intell. Res.(JAIR)*, 26:153–190.

[Jones, 1995] Jones, T. (1995). *Evolutionary algorithms, fitness landscapes and search*. PhD thesis, Citeseer.

[Kikuchi, 1951] Kikuchi, R. (1951). A theory of cooperative phenomena. *Physical Review*, 81(6):988–1003.

[Koller and Friedman, 2009] Koller, D. and Friedman, N. (2009). Probabilistic graphical models: Principles and techniques (adaptive computation and machine learning series).

[Kolmogorov, 2006] Kolmogorov, V. (2006). Convergent tree-reweighted message passing for energy minimization. *IEEE transactions on pattern analysis and machine intelligence*, 28(10):1568–1583.

[Komodakis et al., 2011] Komodakis, N., Paragios, N., and Tziritas, G. (2011). Mrf energy minimization and beyond via dual decomposition. *IEEE transactions on pattern analysis and machine intelligence*, 33(3):531–552.

[Korč et al., 2012] Korč, F., Kolmogorov, V., Lampert, C. H., et al. (2012). Approximating marginals using discrete energy minimization. In *Proc. Workshop Inferning: Interactions between Inference and Learning*. Citeseer.

[Lauritzen, 1996] Lauritzen, S. (1996). *Graphical Models*. Oxford University Press.

[Lin et al., 2016] Lin, P., Neil, M., and Fenton, N. (2016). Region based approximation for high dimensional bayesian network models. *arXiv preprint arXiv:1602.02086*.

[Loh and Wibisono, 2014] Loh, P.-L. and Wibisono, A. (2014). Concavity of reweighted kikuchi approximation. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 3473–3481. Curran Associates, Inc.

[MacKay, 2003] MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

[Minka, 2001a] Minka, T. (2001a). *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Technology, Massachusetts.

[Minka, 2001b] Minka, T. P. (2001b). Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc.

[Nilsson, 1998] Nilsson, D. (1998). An efficient algorithm for finding the M most probable configurations in probabilistic expert systems. *Statistics and Computing*, 2:159–173.

[Pakzad and Anantharam, 2005] Pakzad, P. and Anantharam, V. (2005). Estimation and marginalization using the kikuchi approximation methods. *Neural Computation*, 17(8):1836–1873.

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

[Pelizzola, 2005] Pelizzola, A. (2005). Cluster variation method in statistical physics and probabilistic graphical models. *Journal of Physics A: Mathematical and General*, 38(33):R309.

[Ravikumar et al., 2010] Ravikumar, P., Agarwal, A., and Wainwright, M. J. (2010). Message-passing for graph-structured linear programs: Proximal methods and rounding schemes. *Journal of Machine Learning Research*, 11(Mar):1043–1080.

[Santana, 2005] Santana, R. (2005). Estimation of distribution algorithms with kikuchi approximations. *Evolutionary Computation*, 13(1):67–97.

[Santana et al., 2005] Santana, R., Larranaga, P., and Lozano, J. A. (2005). Properties of kikuchi approximations constructed from clique based decompositions. Technical report, Technical Report EHU-KZAA-IK-2/05, Department of Computer Science and Artificial Intelligence, University of the Basque Country.

[Tatikonda and Jordan, 2002] Tatikonda, S. and Jordan, M. I. (2002). Loopy belief propagation and Gibbs measures. In *Proceedings of the 18th Annual Conference on Uncertainty in Artificial Intelligence (UAI-2002)*, pages 493–500. Morgan Kaufmann Publishers.

[Wainwright et al., 2001] Wainwright, M. J., Jaakkola, T., and Willsky, A. S. (2001). Tree-based reparameterization framework for analysis of belief propagation and related algorithms. Technical Report LIDS P-2510, Laboratory for Information and Decision Systems, MIT.

[Wainwright et al., 2005] Wainwright, M. J., Jaakkola, T. S., and Willsky, A. S. (2005). Map estimation via agreement on trees: message-passing and linear programming. *IEEE transactions on information theory*, 51(11):3697–3717.

[Wainwright and Jordan, 2003] Wainwright, M. J. and Jordan, M. I. (2003). Graphical models, exponential families, and variational inference. Technical Report 649, Department of Statistics, University of California, Berkeley.

[Wainwright and Jordan, 2008] Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305.

[Weiss et al., 2012] Weiss, Y., Yanover, C., and Meltzer, T. (2012). Map estimation, linear programming and belief propagation with convex free energies. *arXiv preprint arXiv:1206.5286*.

[Weller, 2015] Weller, A. (2015). Bethe and related pairwise entropy approximations. In *Uncertainty in Artificial Intelligence (UAI)*.

[Welling et al., 2012] Welling, M., Minka, T. P., and Teh, Y. W. (2012). Structured region graphs: Morphing ep into gbp. *arXiv preprint arXiv:1207.1426*.

[Wormald, 1985] Wormald, N. C. (1985). Counting labelled chordal graphs. *Graphs and combinatorics*, 1(1):193–200.

[Yedidia et al., 2003] Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2003). Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8:236–239.

[Yedidia et al., 2005] Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2005). Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7):2282–2312.

[Yuille, 2001] Yuille, A. (2001). A double-loop algorithm to minimize the Bethe and Kikuchi free energies. *Neural Computation*, 14(6):1691–1722.