Chain Graphical Models
10708, Fall 2020
Pradeep Ravikumar

1 Introduction

So far we have studied UGMs and DGMs each of which are associated with graphs that only contain undirected and directed edges respectively. As we have seen in certain cases, one or the other class is more suitable. What if we could generalize both of these, via a single class that includes both undirected and directed graphs? This is precisely the class of chain graphical models (CGMs). Before setting this up, let us first consider a very popular class of conditional models, called conditional random fields (CRFs).

2 Conditional Random Fields

CRFs as originally formulated were not presented as involving mixed graphs with both undirected and directed edges, so let us set aside the chain graph motivation for now. So far, we have been concerned with joint distributions over a set of variables $X = (X_1, \dots, X_p)$. Let us now consider the conditional variant of this. Let $Y = (Y_1, \dots, Y_d)$ be a set of target or response variables that we wish to predict given some input variables $X = (X_1, \dots, X_p)$. Let $G = (V, E)$ be a "CRF UG" with nodes $V = \mathcal{X} \cup \mathcal{Y}$, but where the undirected edges $E \subseteq (\mathcal{Y} \times \mathcal{Y}) \cup (\mathcal{X} \times \mathcal{Y})$, that is do not include edges between variables within X.

Definition 1 (Conditional Random Field) *Given a CRF UG G as defined above, let G^+ be the extension of G where all variables within $X$are fully connected to each other. Then a conditional distribution $P(Y|X)$ is said to be a conditional random field wrt G if it factors according to G as follows:*

$$P(Y|X) = \frac{1}{Z(X)} \prod_{C \in \mathcal{C}(G^+) \,|\, C \notin \mathcal{X}} \phi_C((Y, X)_C),$$

where $Z(X) = \sum_Y \prod_{C \in \mathcal{C}(G^+) \,|\, C \notin \mathcal{X}} \phi_C((Y, X)_C)$ is the normalization constant.

The key difference between a CRF and a UGM over (X, Y) is that we are not concerned with edges among the variables X, which may have potentially complex dependencies, and moreover might not have a simple or known parametric form for its factors. This allows us to include a varied heterogenous set of input features, without necessarily worrying about their internal dependencies. Instead, we focus only on edges among Y and between Y and

X. The resulting clique factors thus always include some target variable in Y. Additionally, we only use these to specify the conditional distribution $P(Y|X)$ so that the normalization constant $Z(X)$ also depends on the inputs X.

In some treatments of CRFs, a simplified definition is presented, where we have a UG G just over $\mathcal{Y}$, and the clique factors specifying $P(Y|X)$ consist of $\Phi_C(Y_C, X)$ for cliques $C \in G$, so that each clique factor is restricted to target variables in the clique, and *all possible input variables.* In other words, each target variables Y_j is connected to all of the input variables. While any of the CRF factors in the definition we have presented above could be cast in this less restricted form, the converse is not true.

Example 2 (Naive Markov) *Consider the simple case where $d = 1$, and there is a single target variable, and the UG over (X, Y) consists of edges $\{(X_i, Y)\}_{i=1}^p$. Then a strictly positive CRF distribution $P(Y|X)$ is given by:*

$$P(Y|X) = \frac{1}{Z(X)} \exp(\sum_{i=1}^{p} \phi_i(X_i, Y)).$$

A simple instance of such a factor function is $\phi_i(X_i, Y) = w_i X_i Y$, for some weight $w_i \in \mathbb{R}$. When Y is binary, the resulting normalization constant $Z(X)$ is given by:

$$Z(X) = \exp\left(\sum_{i=1}^{p} w_i X_i)\right) + 1,$$

so that the CRF distribution is given by:

$$P(Y = 1|X) = \text{SIGMOID}(\sum_{i=1}^{p} w_i X_i),$$

that is, a logistic regression model.

CRFs as Partially Directed Graphs. It is natural to think of CRFs as a partially directed graph: with directed edges from input variables in X to output variables in Y, while the edges within Y are undirected, and finally where we do not care about edges within X since we only model $P(Y|X)$. Thus, the CRF distribution could be written as $P(Y|\text{PA}_Y)$, where PA_Y are the set of parents — all within X — of the target variables in Y. We can generalize this setup to the more general case of chain graphs.

3 Chain Graphs

Consider a partially directed acyclic graph (PDAG), also known as a chain graph (for reasons which will be clear in the sequel) $G = (V, E)$, where some of the edges are directed, and some are undirected, but there is no directed cycle: there is no path $X_{v_1} - \ldots - X_{v_k} - X_{v_1}$ in G where at least one of the edges is directed; undirected cycles are allowed however. Due to the directed acyclicity, it follows that the set of nodes V can be split into a disjoint partition $\{V_j\}_{j=1}^k$ such that:

- the induced subgraphs $G[V_j]$ for each component V_j has no directed edges
- for any pair of nodes (X, Y), there is a directed edge from $X \to Y$ only when $X \in V_i$ lies in an earlier component than $Y \in V_j$ so that $i < j$.

The components $\{V_j\}_{j=1}^k$ thus form a directed chain, and hence are called chain components, while the PDAG is also called a chain graph.

It can be seen that this generalizes both DAGs and UGs. A DAG is a chain graph, where the chain components consist of individual nodes. While a UG consists of a single chain component.

For each chain component V_j, let $\textsc{pa}_{V_j}$ denote the set of parents of all nodes in V_j. With some overloading of notation, we also denote the moralization of the chain graph G by $\mathcal{M}[G]$: where we fully connect via undirected edges all nodes in $\textsc{pa}_{V_j}$ for each chain component V_j, and then convert all directed edges to undirected ones. This definition can be seen to generalize the moralization of a DAG.

Definition 3 (Chain Graphical Model) *Suppose we have a chain graph G with chain components $\{V_j\}_{j=1}^k$. Let $P(V_j \,|\, \textsc{pa}_{V_j})$ be a CRF over the induced graph over $V_j \cup \textsc{pa}_{V_j}$ of the moralization of G: $G_j := \mathcal{M}[G][V_j \cup \textsc{pa}_{V_j}]$. Thus,*

$$P(X_{V_j} | X_{\textsc{pa}_{V_j}}) = \frac{1}{Z(X_{\textsc{pa}_{V_j}})} \prod_{C \in \mathcal{C}(G_j) \,|\, C \notin \textsc{pa}_{V_j}} \phi_C(X_C),$$

where $Z(X_{\textsc{pa}_{V_j}}) = \sum_{X_{V_j}} \prod_{C \in \mathcal{C}(G_j) \,|\, C \notin \textsc{pa}_{V_j}} \phi_C(X_C)$ is the normalization constant.

Then, a distribution P is said to factor according to the chain graph if it has the form:

$$P(X) = \prod_{j=1}^{k} P(X_{V_j} | X_{\textsc{pa}_{V_j}}).$$

A distribution that factors according to a chain graph G is also called a chain graphical model distribution associated with G.

4 Markov Properties of Chain Graphs: LWF Version

We can generalize the developments of associating UGs and DAGs with conditional independence Markov properties to the encompassing class of chain graphs. It turns out however that there are many ways to do that, all of which "marginalize" to the correct set of Markov properties for UGs and DAGs. We will start out by discussing the most common extension, also called LWF Markov properties, due to the authors Lauritzen, Wermuth and Frydenberg.

We first extend some standard graph theoretic terminology to chain graphs. As with DAGs, we have even for CGs:

$$\begin{aligned}\textsc{pa}_X &= \{Y \in V \text{ s.t. } Y \to X \in E\},\\ \textsc{child}_X &= \{Y \in V \text{ s.t. } X \to Y \in E\}.\end{aligned}$$

Also, as with UGs, $\textsc{nbrs}_X = \{Y \in V \text{ s.t. } Y - X \in E\}$. The notions of ancestor and descendant are a bit more subtle. $\textsc{ances}_X$ consists all nodes $Y \in V$ s.t. there exists a path consisting solely of undirected and directed edges from G from Y to X. Similarly, $\textsc{desc}_X$ consists all nodes $Y \in V$ s.t. there exists a path consisting solely of undirecte and at least one directed edges from G from X to Y.

For any node X, we define its boundary $\textsc{boundary}_X = \textsc{pa}_X \cup \textsc{nbrs}_X$.

Definition 4 (Pairwise Markov properties) *Given a chain graph G, we define the set of pairwise Markov independencies as the set:*

$$\mathbb{I}_p(G) = \{X \perp\!\!\!\perp Y \,|\, \textsc{non-desc}_X - \{X, Y\} \;:\; (X, Y) \notin E(G),\, Y \in \textsc{non-desc}_X\}.$$

Note that for a UG, $\textsc{non-desc}_X = V$, so that this exactly corresponds to a pairwise Markov property for UGs. It also matches as is the pairwise Markov property for DAGs. Moving onto the broader set of local Markov properties, we have the following.

Definition 5 (Local Markov properties) *Given a chain graph G, we define the set of local Markov independencies as the set:*

$$\mathbb{I}_\ell(G) = \{X \perp\!\!\!\perp \textsc{non-desc}_X - \textsc{boundary}_X \,|\, \textsc{boundary}_X\}.$$

Recall that for DAGs, $\textsc{boundary}_X = \textsc{pa}_X$, so this exactly corresponds to the local Markov properties in DAGs. While for UGs, $\textsc{boundary}_X = \textsc{nbrs}_X$, and $\textsc{non-desc}_X = V$, which exactly corresponds to the local Markov property in UGs.

Let us now consider the counterpart of global Markov properties in chain graphs.

Definition 6 (c-separation) *Given a chain graph G, consider any three disjoint sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$. Let $U = \mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z}$. We then say that $\mathbf{X}$ is c-separated from $\mathbf{Y}$ given $\mathbf{Z}$, also denoted as* $\textsc{csep}(\mathbf{X}, \mathbf{Y} \,|\, \mathbf{Z})$*, if $\mathbf{X}$ is separated from $\mathbf{Y}$ given $\mathbf{Z}$ in the UG $\mathcal{M}[G[U \cup \textsc{ances}_U]]$, which is the chain-graph-moralization of the induced subgraph over nodes in U and their ancestors.*

It can be seen that *c*-separation reduces to the notion of *d*-separation for DAGs. And since the set of ancestors of any node in a UG is simply the rest of the graph, this also reduces to simple graph separation for UGs. Armed with this terminology, we can then define the set of global Markov properties for CGs.

Definition 7 (Global Markov properties) *Given a chain graph G, we define the set of global Markov independencies as the set:*

$$\mathbb{I}(G) = \{\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z} \text{ s.t. } \textsc{csep}(\mathbf{X}, \mathbf{Y} \,|\, \mathbf{Z}) \; \forall \text{ disjoint } \mathbf{X}, \mathbf{Y}, \mathbf{Z} \in V\}.$$

As with UGs, we have the following proposition:

Proposition 8

$$\mathbb{I}_p(G) \subseteq I_\ell(G) \subseteq I(G).$$

And moreover that for positive distributions, we have an equivalence between distributions that satisfy any of these properties, as well as distributions that factor according to G.

Proposition 9 *For any positive distribution P, and a chain graph G, the following statements are equivalent:*

1. *P factors according to G*
2. *P satisfies cond. independencies in $\mathbb{I}_p(G)$*
3. *P satisfies cond. independencies in $\mathbb{I}_\ell(G)$*
4. *P satisfies cond. independencies in $I(G)$*

5 Markov Properties of Chain Graphs: AMP Version

Consider the graph in Figure 1. The chain components are $\{1, 2\}$, and $\{3, 4\}$. The LWF Markov properties would entail that:

$$1 \perp\!\!\!\perp 4 \,|\, \{2, 3\}$$
$$2 \perp\!\!\!\perp 3 \,|\, \{1, 4\}.$$

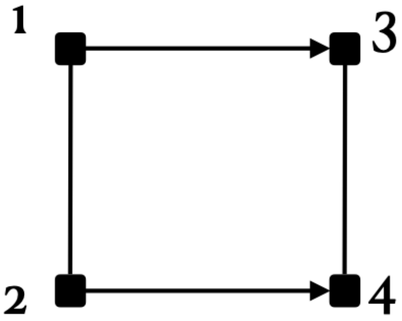

Figure 1: A Chain Graph G

These might seem like natural enough conditional independencies entailed by the graph. But consider a very natural *generative process*, also known as a structural equation model (SEM) suggested by the graph:

$$\begin{aligned} X_1 &= \epsilon_1 \\ X_2 &= \epsilon_2 \\ X_3 &= b_{31}X_1 + \epsilon_3 \\ X_4 &= b_{42}X_2 + \epsilon_4, \end{aligned}$$

where the noise vectors in the two chain components are independent of each other so that $(\epsilon_1, \epsilon_2) \perp\!\!\!\perp (\epsilon_3, \epsilon_4)$, but where the noise variables within a chain component could be dependent: so that ϵ_1 and ϵ_2 could be dependent (e.g. a non spherical bivariate Gaussian), as could ϵ_3 and ϵ_4. It can be seen that when ϵ_1 and ϵ_2 are dependent, as are ϵ_3 and ϵ_4, then the LWF properties need no longer hold. Instead, for the SEM above, it can be seen that we instead have a different set of conditional independencies:

$$\begin{aligned} 1 \perp\!\!\!\perp 4 \,|\, 2 \\ 2 \perp\!\!\!\perp 3 \,|\, 1, \end{aligned}$$

which seems more intuitive when reading the graph as a generative process. These latter properties are also called AMP Markov properties, after the authors Andersson, Madigan and Perlman, though the authors, perhaps jokingly, suggest it is named after Alternative Markov Property.

For more intuition of the difference between the LWF and the AMP properties, it is instructive to consider the case were (X_1, X_2, X_3, X_4) are multivariate normal with mean $\mathbf{0}$

and covariance matrix $\Sigma \in \mathbb{R}^{4\times 4}$. The conditional distribution of (X_3, X_4) given (X_1, X_2) is given by:

$$\begin{pmatrix} X_3 \\ X_4 \end{pmatrix} | X_1, X_2 \sim \mathcal{N}(B \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \Lambda),$$

for some matrices B, Λ that depend on Σ. Letting $B \equiv \left(\begin{smallmatrix} B_{31} & B_{32} \\ B_{41} & B_{42} \end{smallmatrix}\right)$, The AMP conditions above are equivalent to the very interpretable:

$$B_{32} = B_{41} = 0.$$

In particular, note that the conditional mean $\mathbb{E}[X_3 | X_1, X_2] = B_{31}X_1 + B_{32}X_2 = B_{31}X_1$, which does not depend on X_2 when $B_{32} = 0$.

The LWF conditions on the other hand work with the matrix $\Gamma \equiv \left(\begin{smallmatrix} \Gamma_{31} & \Gamma_{32} \\ \Gamma_{41} & \Gamma_{42} \end{smallmatrix}\right) = \Lambda^{-1}B$, and are equivalent to the much less interpretable:

$$\Gamma_{32} = \Gamma_{41} = 0.$$

The advantage of Γ is that it is a part of the natural or canonical parameterization of the conditional distribution of (X_3, X_4) given (X_1, X_2):

$$P((X_3, X_4) \,|\, X_1, X_2) \propto \exp(-\Gamma \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} - 1/2 \begin{pmatrix} X_3 \\ X_4 \end{pmatrix}^T \Lambda^{-1} \begin{pmatrix} X_3 \\ X_4 \end{pmatrix}).$$

In contrast to the very intuitive AMP consequence of the conditional mean of X_3 being independent of X_2 when $B_{32} = 0$, here it might seem much less intuitive to ask for the canonical parameters $\Gamma_{32} = 0$. But on the other hand, the corresponding canonical parameters being zero entails that the factorized form of the conditional distribution is more intuitive:

$$P((X_3, X_4) \,|\, X_1, X_2) \propto \phi_{34}(X_3, X_4)\phi_{13}(X_1, X_3)\phi_{24}(X_2, X_4),$$

which seems more natural.

Thus whether LWF or AMP is more natural depends on the form of our parameterization. With SEMs, AMP properties might be more intuitive, while with clique factorized forms, the LWF properties seem more intuitive.

5.1 Characterizations of AMP Properties

So how do we characterize the AMP properties? We first need the notion of a flag or a double-flag:

Definition 10 (Flag) *A 3-tuple of nodes (X, Y, Z) is a flag wrt a chain graph G if the following edges belong to $E(G)$:*

- $X \to Y - Z$
- $X - Y \leftarrow Z$
- $X \to Y \leftarrow Z$

Definition 11 (Double Flag) *A 4-tuple of nodes (X, Y, Z, U) is a double-flag wrt a chain graph G if the following edges belong to $E(G)$:*

- $X \to Y - Z$
- $U \to Z - Y$

A flag (X, Y, Z) is augmented by adding the edge (X, Z). A double flag (X, Y, Z, U) is augmented by adding the edges $(X, Z), (Y, U), (X, U)$. We say that $\mathcal{A}[G]$ is the augmentation of a chain graph G if all its flags and double-flags are augmented. Note that with DAGs, the only flag possible is an immorality, in which case augmentation exactly corresponds to DAG-moralization.

Definition 12 (AMP-separation) *Given a chain graph G, consider any three disjoint sets $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$. Let $U = \mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z}$. We then say that $\mathbf{X}$ is AMP-separated from $\mathbf{Y}$ given $\mathbf{Z}$, or that $\mathrm{AMP}(\mathbf{X}, \mathbf{Y} \,|\, \mathbf{Z})$ holds if $\mathbf{X}$ is separated from $\mathbf{Y}$ given $\mathbf{Z}$ in the UG $\mathcal{A}[G[U \cup \textsc{ances}_U]]$, which is the chain-graph-augmentation of the induced subgraph over nodes in U and their ancestors.*

It can be seen that AMP-separation reduces to the notion of d-separation for DAGs, since for DAGs augmentation exactly corresponds to moralization. And since the set of ancestors of any node in a UG is simply the rest of the graph, this also reduces to simple graph separation for UGs. We thus see that AMP-separation is a completely distinct notion of chain graph separation from LWF c-separation (since chain-graph augmentation is distinct from chain-graph moralization), but both of these reduce to the same notion in either UGs or DAGs. Armed with this terminology, we can then define the set of AMP-global Markov properties for CGs.

Definition 13 (AMP Global Markov properties) *Given a chain graph G, we define the set of AMP global Markov independencies as the set:*

$$\mathbb{I}_{\mathrm{AMP}}(G) = \{\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\, \mathbf{Z} \ \textit{s.t.}\ \mathrm{AMP}(\mathbf{X}, \mathbf{Y} \,|\, \mathbf{Z}) \forall \textit{ disjoint } \mathbf{X}, \mathbf{Y}, \mathbf{Z} \in V\}.$$

The following proposition follows naturally from the definition of augmentation.

Proposition 14 *If the chain graph G has no flags other than immoralities, then $\mathbb{I}_{\mathrm{LWF}}(G) = \mathbb{I}_{\mathrm{AMP}}(G)$.*

To get some more intuition about the difference between LWF and AMP, and indeed, even about LWF itself, let us consider the following equivalent characterization of LWF Markov properties when restricted to positive distributions. For any graph G with chain components $\{V_j\}_{j=1}^{p}$, consider the DAG $G_D = (V_D, E_D)$ where the nodes correspond to the chain components $V_D = \{V_j\}_{j=1}^{p}$, and the edges $E_D = \{(V_j, V_k) \text{ s.t. } X \rightarrow Y \in G, X \in V_j, Y \in V_k\}$. And the UGs $G_j = (V_j, E_j)$ corresponding to the individual chain components. We can then express the LWF Markov properties entirely in terms of the DAG Markov properties over the DAG G_D, the UG Markov properties over the UGs $\{G_j\}$, and a connective Markov property that involves both directed and undirected edges.

Definition 15 (Block-Recursive Characterization of LWF) *Any positive distribution P satisfies the LWF global Markov properties iff it satisfies the following three properties:*

- $\forall j \in [k]$, $V_j \perp\!\!\!\perp \textsc{non-desc}_D(V_j) - \textsc{pa}_D(V_j) \,|\, \textsc{pa}_D(V_j)$ *i.e. P satisfies the local Markov properties with respect to the DAG G_D*
- $\forall j \in [k]$, *$P(V_j \,|\, \textsc{pa}_D(V_j))$ satisfies global Markov properties with respect to G_j*
- $\forall j \in [k]$, $\forall U \subseteq V_j, U \perp\!\!\!\perp \textsc{pa}_D(V_j) - \textsc{pa}_G(U) \,|\, \textsc{pa}_G(U) \cup (V_j - U)$.

The first two properties are very natural, whereas the last is perhaps less so. It states that any subset of a chain component is independent of its non-parent nodes in the parent chain components conditioned on its parent nodes AND rest of variables in its chain component.

We can consider the analogous construction for AMP properties.

Definition 16 (Block-Recursive Characterization of AMP) *Any positive distribution P satisfies the AMP global Markov properties iff it satisfies the following three properties:*

- $\forall j \in [k]$, $V_j \perp\!\!\!\perp \textsc{non-desc}_D(V_j) - \textsc{pa}_D(V_j) \,|\, \textsc{pa}_D(V_j)$ *i.e. P satisfies the local Markov properties with respect to the DAG G_D*
- $\forall j \in [k]$, *$P(V_j \,|\, \textsc{pa}_D(V_j))$ satisfies global Markov properties with respect to G_j*
- $\forall j \in [k]$, $\forall U \subseteq V_j, U \perp\!\!\!\perp \textsc{pa}_D(V_j) - \textsc{pa}_G(U) \,|\, \textsc{pa}_G(U)$.

The AMP characterization only differs in the last condition. It states that any subset of a chain component is independent of its non-parent nodes in the parent chain components conditioned on its parent nodes. It no longer requires conditioning on other variables within its chain component. This might perhaps seem more natural.

Nonetheless, the most common Markov property characterization is the set of LWF properties. This is because as noted earlier, LWF properties are more natural than AMP when we work with distributions in clique factorized form, and these are the most popular forms in use within statistical machine learning. But causal graphical models are getting more popular in ML, and there we work with the SEM factorization, so that the AMP properties might well make a comeback.