Chapter 2

Background

In this chapter we provide a brief overview of the foundations on which this thesis builds, particularly probabilistic graphical models, exponential family distributions, hidden Markov models, and Bayesian nonparametric models constructed using the Dirichlet process.

■ 2.1 Graphical models

In this section we overview the key definitions and results for directed and undirected probabilistic graphical models, which we use both for defining models and constructing algorithms in this thesis. For a more thorough treatment of probabilistic graphical models, see Koller and Friedman [65].

■ 2.1.1 Directed graphical models

Directed graphical models, also called Bayes nets, naturally encode generative model parameterizations, where a model is specified via a sequence of conditional distributions. They are particularly useful for the hierarchical Bayesian models and algorithms developed in this thesis.

First, we give a definition of directed graphs and a notion of directed separation of nodes. Next, we connect these definitions to conditional independence structure for collections of random variables and factorization of joint densities.

Definition 2.1.1 (Directed graph)**.** *For some $n \in \mathbb{N}$, a* directed graph *on n nodes is a pair (V, E) where $V = [n] \triangleq \{1, 2, \ldots, n\}$ and $E \subseteq (V \times V) \setminus \{(i, i) : i \in V\}$. We call the elements of V the (labeled)* nodes *or vertices and the elements of E the* edges*, and we say $(i, j) \in E$ is an edge from i to j.*

Given a graph (V, E), for distinct $i, j \in V$ we write $i \to j$ or $j \leftarrow i$ if $(i, j) \in E$ and write $i - j$ if $(i, j) \in E$ or $(j, i) \in E$. We say there is a *directed path* from i_1 to i_n of length $n - 1$ if for some $i_2, i_3, \ldots, i_{n-1} \in V$ we have $i_1 \to i_2 \to \cdots \to i_n$, and an *undirected path* if we have $i_1 - i_2 - \cdots - i_n$. We say node j is a *descendant* of node i

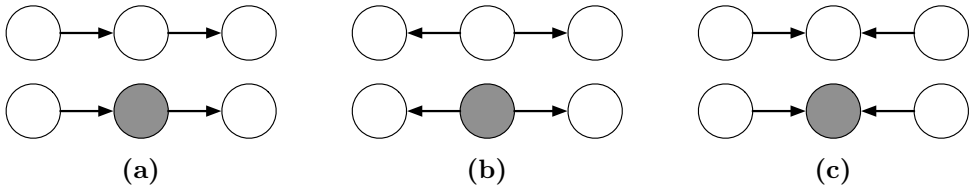

Figure 2.1: An illustration of the cases in Definition 2.1.2. Shaded nodes are in the set C.

if there is a directed path from i to j, and we say j is a *child* of i if there is a directed path of length 1. Similarly, we say i is an *ancestor* of j if j is a descendant of i, and we say i is a *parent* of j if j is a child of i. We use $\pi_G(i)$ to denote the set of parents of node i and $c_G(i)$ to denote its children.

We say a directed graph is *cyclic* if there is a node $i \in V$ that is its own ancestor, and we say a graph is *acyclic* if it is not cyclic. For directed graphical models, and all of the directed graphs in this thesis, we use *directed acyclic graphs* (DAGs).

Using these notions we can define the main idea of *directed separation*.

Definition 2.1.2 (Blocked/unblocked triples)**.** *Given a DAG $G = (V, E)$, let $a - b - c$ be an undirected path with $a, b, c \in V$, and let $C \subset V$ be a subset of nodes with $C \cap \{a, c\} = \varnothing$. We call $a - b - c$ a* triple*, and we say it is* blocked by C *in two cases:*

1. *if the structure is not $a \to b \leftarrow c$, then $b \in C$*
2. *if the structure is $a \to b \leftarrow c$, and for all descendants $b' \in V$ of b we have $b' \notin C$.*

We say a triple is unblocked by C *if it is not blocked by C.*

We illustrate the cases in Definition 2.1.2 in Figure 2.1, which shows six triples of nodes, where nodes in the set C are shaded. In each of (a) and (b), the top triple is unblocked while the bottom triple is blocked, corresponding to case 1 in the definition. However, in (c) the reverse is true: the top triple is blocked while the bottom triple is unblocked, corresponding to case 2 in the definition.

Definition 2.1.3 (Blocked/unblocked path)**.** *Given a DAG $G = (V, E)$ and a set $C \subset V$, let $i_1 - i_2 - \cdots - i_n$ be a path with $C \cap \{i_1, i_n\} = \varnothing$. We call the path* unblocked by C *if every triple in the path is unblocked by C. We call the path* blocked by C *if it is not unblocked.*

Note that Koller and Friedman [65] uses the term *active trail* for our definition of unblocked path.

Definition 2.1.4 (d-separation)**.** *Given a DAG $G = (V, E)$, for distinct $i, j \in V$ and a subset $C \subset V$ with $C \cap \{i, j\} = \varnothing$, we say i and j are* d-separated in G by C

if there is no undirected path between i and j that is unblocked by C, and we write $\text{d-sep}_G(i, j|C)$. *Further, for disjoint subsets $A, B, C \subset V$ with A and B nonempty we write* $\text{d-sep}_G(A, B|C)$ *if we have* $\text{d-sep}_G(i, j|C)$ *for all $i \in A$ and $j \in B$.*

In words, $i, j \in V$ may not be d-separated in G given C if there exists an undirected path between i and j in G. However, the path must be unblocked, where if a node on the path belongs to C it generally blocks the path except when there is a "V" structure $a \to b \leftarrow c$ on the path, in which case b blocks the path unless it or one of its descendants is in C. This special rule is useful when defining probabilistic structure in terms of the graph because it models how independent random variables can become dependent when they are competing explanations for the same observation.

Next, we give a definition of conditional independence structure in collections of random variables that uses graphical d-separation.

Definition 2.1.5 (Markovianity on directed graphs)**.** *Given a DAG $G = (V, E)$ and a collection of random variables $X = \{X_i : i \in V\}$ indexed by labeled nodes in the graph, we say X is* Markov on G *if for disjoint subsets $A, B, C \subset V$ we have*

$$\text{d-sep}_G(A, B|C) \implies X_A \perp\!\!\!\perp X_B | X_C \tag{2.1.1}$$

where for $S \subseteq V$ we define $X_S \triangleq \{X_i : i \in S\}$.

Note that this definition does not require that the graph capture all of the conditional independencies present in the collection of random variables. Indeed, a collection of random variables can be Markov on many distinct graphs, and every collection is Markov on the complete graph. Graphs that capture more structure in the collection of random variables are generally more useful.

Conditional independence structure can be used in designing inference algorithms, and a graphical representation can make clear the appropriate notion of local information when designing an algorithm with local updates. A particularly useful notion of local information is captured by the *Markov blanket*.

Definition 2.1.6 (Directed Markov blanket)**.** *Given a DAG $G = (V, E)$, the* Markov blanket *for node $i \in V$, denoted $\text{MB}_G(i)$, is the set of its parents, children, and childrens' parents:*

$$\text{MB}_G(i) \triangleq \{j \in V : j \to i\} \cup \{j \in V : i \to j\} \cup \{j \in V : \exists k \,.\, i \to k \leftarrow j\}. \tag{2.1.2}$$

The Markov blanket for a set of nodes $A \subseteq V$ contains the Markov blankets for all nodes in A except the nodes in A itself:

$$\text{MB}_G(A) \triangleq \bigcup_{i \in A} \text{MB}_G(i) \setminus A. \tag{2.1.3}$$

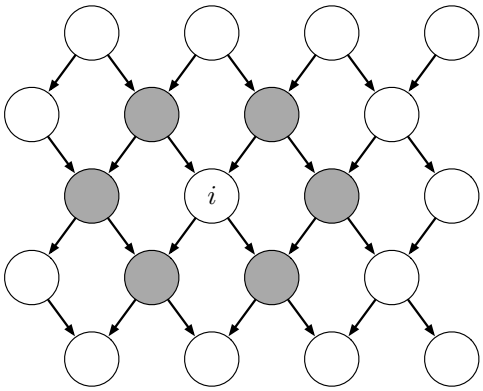

Figure 2.2: An illustration of the directed Markov blanket defined in Definition 2.1.6. Nodes in the Markov blanket of node i are shaded gray.

We illustrate Definition 2.1.6 in Figure 2.2. The nodes in the Markov blanket of node i are shaded gray.

Proposition 2.1.7. *Given a collection of random variables $\{X_i : i \in V\}$ that is Markov with respect to a DAG $G = (V, E)$, we have*

$$X_i \perp\!\!\!\perp X_S | X_{\mathrm{MB}_G(i)} \tag{2.1.4}$$

where $S \triangleq V \setminus (\mathrm{MB}_G(i) \cup \{i\})$.

Proof. By conditioning on the parents of node i, all paths of the form $a \to b \to i$ are blocked. By conditioning on its children, all paths of the form $i \to b \to c$ are blocked. By conditioning on the childrens' parents, all paths of the form $i \to b \leftarrow c$, which may have been unblocked by conditioning on b or one of its descendants via case 2 of Definition 2.1.4, are blocked. □

Another common and convenient notion of probabilistic graphical structure is a density's factorization with respect to a DAG.

Definition 2.1.8 (Factoring on directed graphs)**.** *Given a DAG $G = (V, E)$ on n nodes and a collection of random variables $X = \{X_i : i \in V\}$ with density p_X with respect to some base measure, we say p_X* factorizes according to *G if we can write*

$$p_X(x_1, \ldots, x_n) = \prod_{i \in V} p(x_i | x_{\pi_G(i)}) \tag{2.1.5}$$

where $\pi_G(i) \triangleq \{j \in V : j \to i\}$ denotes the set of parents of node i in G.

Theorem 2.1.9. *A collection of random variables $\{X_i : i \in V\}$ with a joint density (with respect to some base measure) is Markov on a DAG G if and only if the joint density factorizes as in Eq. (2.1.5).*

Proof. The proof is straightforward. In Koller and Friedman [65], Theorem 3.1 shows Markovianity implies the densities factor and Theorem 3.2 shows the reverse. □

■ 2.1.2 Undirected graphical models

Undirected graphical models, also called Markov random fields, do not easily encode generative model specifications. However, they can be more useful for encoding soft constraints or local partial correlations. We use an undirected graphical model perspective in our analysis of Hogwild Gibbs Sampling in Chapter 7.

As with directed graphical models, we first define undirected graphs and a notion of separation of nodes, then give definitions that link the graphical structure to both conditional independence structure in a collection of random variables and factorization structure in the joint density for those variables.

Definition 2.1.10 (Undirected graph)**.** *For some $n \in \mathbb{N}$, an* undirected graph *on n nodes is a pair (V, E) where $V = [n]$ and $E \subseteq \{\{i, j\} : i, j \in V,\ i \neq j\}$.*

Analogous to the definition in the previous section, there is a natural notion of an undirected path between nodes. Given a graph (V, E), for distinct $i, j \in V$ we write $i - j$ if $\{i, j\} \in E$, and we say there is an (undirected) *path* from i_1 to i_n of length $n-1$ if for some $i_2, i_3, \ldots, i_{n-1} \in V$ we have $i_1 - i_2 - \cdots - i_n$. We say i is a *neighbor* of j if $\{i, j\} \in E$ and denote the set of neighbors of node i as $n_G(i) \triangleq \{j \in V : \{i, j\} \in E\}$. We say a pair of nodes is *connected* if there exists an (undirected) path from i to j.

The notion of undirected separation and corresponding notion of Markovianity on undirected graphs is simpler than those for directed graphs.

Definition 2.1.11 (Undirected separation)**.** *Given an undirected graph $G = (V, E)$, for distinct $i, j \in V$ and a subset $C \subset V$ with $C \cap \{i, j\} = \varnothing$, we say i and j are* separated in G given C *and write $\mathrm{sep}_G(i, j|C)$ if there is no path from i to j that avoids C. For disjoint subsets $A, B, C \subset V$ with A and B nonempty we write $\mathrm{sep}_G(A, B|C)$ if we have $\mathrm{sep}_G(i, j|C)$ for all $i \in A$ and $j \in B$.*

Definition 2.1.12 (Markovianity on undirected graphs)**.** *Given an undirected graph $G = (V, E)$ and a collection of random variables $X = \{X_i : i \in V\}$ indexed by labeled nodes in the graph, we say X is* Markov on G *if for disjoint subsets $A, B, C \subset V$ we have*

$$\mathrm{sep}_G(A, B|C) \implies X_A \perp\!\!\!\perp X_B | X_C. \tag{2.1.6}$$

As with Markovianity with respect to directed graphs, an undirected graph may not encode all the conditional independence statements possible for a given collection of random variables.

The notion of Markov blanket for an undirected graph is also simpler.

Definition 2.1.13 (Undirected Markov blanket)**.** *Given an undirected graph $G = (V, E)$, the* Markov blanket *for each node $i \in V$, denoted $\mathrm{MB}_G(i)$, is the set of neighbors of node i:*

$$\mathrm{MB}_G(i) \triangleq n_G(i) = \{j \in V : \{i, j\} \in E\}. \tag{2.1.7}$$

The Markov blanket for a set of nodes $A \subseteq V$ is the set of all neighbors to nodes in A excluding those in A:

$$\mathrm{MB}_G(A) \triangleq \bigcup_{i \in A} \mathrm{MB}_G(i) \setminus A. \tag{2.1.8}$$

We can use this definition for an undirected analog of Proposition 2.1.7.

Proposition 2.1.14. *Given a collection of random variables $\{X_i : i \in V\}$ that is Markov with respect to an undirected graph $G = (V, E)$, we have*

$$X_i \perp\!\!\!\perp X_S | X_{\mathrm{MB}_G(i)} \tag{2.1.9}$$

where $S \triangleq V \setminus (\mathrm{MB}_G(i) \cup \{i\})$.

Proof. Because all the neighbors of i are in $\mathrm{MB}_G(i)$, for any $j \in S$ there can be no undirected path from j to i that avoids $\mathrm{MB}_G(i)$. □

We can also define a density factorization with respect to an undirected graph, though we must first define the notion of a clique. A *clique* in a graph (V, E) is a nonempty subset of fully-connected nodes; that is, a nonempty set $C \subseteq V$ is a clique if for every distinct $i, j \in C$ we have $\{i, j\} \in E$.

Definition 2.1.15 (Factoring on undirected graphs)**.** *Given an undirected graph $G = (V, E)$ and a collection of random variables $X = \{X_i : i \in V\}$ with density p_X with respect to some base measure, we say p_X* factorizes according to *G if we can write*

$$p_X(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C) \tag{2.1.10}$$

for a collection of cliques $\mathcal{C}$ of G and nonnegative potentials *or* factors *$\{\psi_C : C \in \mathcal{C}\}$ indexed by those cliques, where*

$$Z \triangleq \int \prod_{C \in \mathcal{C}} \psi_C(x_C) \nu(dx) \tag{2.1.11}$$

is a normalizing constant.

Note that cliques typically overlap; that is, we may have $C_i \cap C_j \neq \varnothing$ for distinct $C_i, C_j \in \mathcal{C}$. To remove many possible redundancies, without loss of generality one can assume that $\mathcal{C}$ includes only *maximal cliques*, where a maximal clique cannot have any other node added to it and remain a (fully-connected) clique.

The correspondence between a collection of random variables being Markov on an undirected graph and its joint density factorizing as Eq. (2.1.10) is not quite as simple as that for directed graphs because deterministic relationships among the random variables can prevent factoring the density, as shown in the next example.

Example 2.1.16. *Using Example 4.4 from Koller and Friedman [65], consider four binary random variables $X = \{X_i : i = 1, 2, 3, 4\}$ with a PMF that takes value $1/8$ on the configurations of (x_1, x_2, x_3, x_4) given by*

$$\begin{array}{cccc} (0,0,0,0) & (1,0,0,0) & (1,1,0,0) & (1,1,1,0) \\ (0,0,0,1) & (0,0,1,1) & (0,1,1,1) & (1,1,1,1) \end{array}$$

and 0 elsewhere. Then X is Markov on the graph $1 - 2 - 3 - 4 - 1$ but the density cannot be factorized into pairwise potentials.

This issue cannot arise when we restrict our attention to strictly positive densities.

Theorem 2.1.17. *Given a collection of random variables $X = \{X_i : i \in V\}$ with a joint density p_X (with respect to some base measure) and an undirected graph G, we have*

1. *If p_X factorizes according to G, then X is Markov on G.*
2. *If X is Markov on G and p_X is strictly positive, then p_X factors according to G.*

Proof. The proof for 1 is straightforward and is given as the proof of Theorem 4.1 in Koller and Friedman [65]. The proof for 2 is the proof of the Hammersley-Clifford theorem, given as Theorem 4.8 in Koller and Friedman [65]. □

■ 2.1.3 Exact inference and graph structure

Given some specification of a probability distribution, *inference* for that distribution means computing quantities of interest such as marginals, conditionals, or expectations. In the graphical model framework we can be precise about the computations required to perform inference and their complexity, as we overview in this subsection. For concreteness, we focus on undirected models with densities.

Consider an undirected graphical model specified by an undirected graph $G = (V, E)$ and a set of potentials $\{\psi_C : C \in \mathcal{C}\}$ on a set $\mathcal{C}$ of cliques of G. The joint density is proportional to the product of the clique potentials $p(x) \propto \prod_{C\in\mathcal{C}} \psi_C(x_C)$. We can represent an arbitrary inference query in terms of a subset of nodes on which we condition, a subset of nodes which we marginalize out, and a subset over which we want to represent the resulting density; that is, we partition the set of nodes V into three disjoint (possibly empty) subsets A_1, A_2, A_3 with $A_1 \cup A_2 \cup A_3 = V$ and write

$$p(x_{A_1}|x_{A_3}) = \int p(x_{A_1}, x_{A_2}|x_{A_3})\nu(dx_{A_2}) = \frac{\int \prod_{C\in\mathcal{C}} \psi_C(x_C)\nu(dx_{A_2})}{\int \prod_{C\in\mathcal{C}} \psi_C(x_C)\nu(dx_{A_1\cup A_2})}. \tag{2.1.12}$$

Therefore to compute an arbitrary inference query we need only to compute integrals of products of the factors in the density. To simplify notation, we write such computations in the form

$$\int \prod_{C\in\mathcal{C}} \psi_C(x_C)\nu(dx_A) \tag{2.1.13}$$

for some subset $A \subseteq V$.

Graph structure affects how we can organize a computation of the form (2.1.13) and thus its computational complexity. Consider the special case of integrating out a single variable x_j by partitioning the set of cliques into those which contain node j and those which do not, $\mathcal{C}_j \triangleq \{C \in \mathcal{C} : j \in C\}$ and $\mathcal{C}_{\setminus j} \triangleq \{C \in \mathcal{C} : j \notin C\}$, and writing

$$\int \prod_{C\in\mathcal{C}} \psi_C(x_C)\nu(dx_j) = \prod_{C_{\setminus j}\in\mathcal{C}_{\setminus j}} \psi_{C_{\setminus j}}(x_{C_{\setminus j}}) \int \prod_{C_j\in\mathcal{C}_j} \psi_{C_j}(x_{C_j})\nu(dx_j) \tag{2.1.14}$$

$$= \prod_{C_{\setminus j}\in\mathcal{C}_{\setminus j}} \psi_{C_{\setminus j}}(x_{C_{\setminus j}})\psi_B(x_B) \tag{2.1.15}$$

where $B \triangleq \{i \in C_j : C_j \in \mathcal{C}_j\} \setminus \{j\}$ is the set of all indices that share a clique with node j in G and ψ_B is a new factor on the clique B resulting from the integral over x_j. Thus as a result of the integral there is an *induced graph* on $V \setminus \{j\}$ formed by *eliminating* j by fully connecting its neighbors and deleting it from the graph.

When integrating over multiple variables, the process repeats: given an *elimination order*, nodes are eliminated one by one from the graph, and each elimination introduces a new clique in the graph and a corresponding new potential term in the density over the remaining variables. The computational complexity of the process is determined by the size of the largest clique encountered. In the case of PMFs with finite support, the number of entries in the table encoding the new potential formed in (2.1.15) is typically exponential in the size of the new clique; in the case of PDFs, the complexity

depends on the complexity of integrating over factors on cliques, where the clique size determines the number of dimensions in the domain of the integrand, and the complexity of representing the result. The optimal elimination order is itself NP hard to find for general graphs.

Some classes of graphs have straightforward elimination orderings that avoid the growth in clique size, and inference in distributions that are Markov on such graphs avoids the corresponding growth in complexity.

Definition 2.1.18 (Tree graph)**.** *A directed or undirected graph $G = (V, E)$ is a* tree *if for each pair distinct nodes $i, j \in V$ there is a unique undirected path from i to j in G.*

In an undirected tree G, we refer to the nodes i with only single neighbors $|n_G(i)| = 1$ as *leaves*. In a directed tree, we refer to the nodes with no children as leaves.

Proposition 2.1.19 (Markov on trees)**.** *A density p that factors on an undirected tree (V, E) can be written in terms of pairwise factors $\{\psi_{ij} : \{i, j\} \in E\}$ as*

$$p(x) = \frac{1}{Z} \prod_{\{i,j\} \in E} \psi_{ij}(x_i, x_j). \tag{2.1.16}$$

Proof. All of the maximal cliques in an undirected tree are of size at most 2. □

Note that because the edges are undirected, ψ_{ij} and ψ_{ji} denote the same object.

Given a density that factors according to an undirected tree specified in terms of its pairwise factors $\{\psi_{ij} : \{i, j\} \in E\}$ we can convert it to a directed specification by local normalization, i.e. by choosing a direction for each edge and computing

$$p(x_i | x_j) = \frac{\psi_{ij}(x_i, x_j)}{\int \psi_{ij}(x_i, x_{j'}) \nu(dx_{j'})}. \tag{2.1.17}$$

Note that a density that factors on a directed tree may not in general be written purely in terms of conditional probability factors that depend only on pairs of nodes, as shown in the next example.

Example 2.1.20. *A density that factors according to the directed tree $G = (V, E)$ with $V = \{1, 2, 3\}$ and edges $1 \to 2 \leftarrow 3$ may include a factor $p(x_2|x_1, x_3)$. Such a density is not Markov on the undirected tree $G' = (V', E')$ with $V' = V$ and edges $1 - 2 - 3$, and instead only factors on the complete undirected graph with edges $1 - 2 - 3 - 1$.*

Directed trees in which each node has at most one parent avoid this issue, and we can convert freely between directed and undirected tree parameterizations for such models.

Proposition 2.1.21. *A density p that factors on a directed tree $G = (V, E)$ in which each node only has one parent, i.e. $|\pi_G(i)| \leq 1$ for $i \in V$, also factors with respect to the undirected tree $G' = (V', E')$ formed by dropping the directions on the edges of G, i.e. $V' = V$ and $E' = \{\{i, j\} : (i, j) \in E\}$*

Proof. Set $\psi_{ij}(x_i, x_j) = p(x_i|x_j)$. □

With undirected trees and directed trees in which each node has at most one parent we can perform elimination without introducing larger factors by using an elimination order that recursively eliminates leaves. Furthermore, the corresponding partial sums for all such elimination orderings can be computed simultaneously with an efficient dynamic programming algorithm, as shown in the next theorem.

For an undirected graph $G = (V, E)$ and a subset of nodes $A \subseteq V$, we define $G \setminus A$ to be the graph formed by deleting the nodes A and the edges incident on nodes in A, i.e. the graph (V', E') where $V' = V \setminus A$ and $E' = \{\{i, j\} \in E : i, j \notin A\}$. We say a pair of nodes is *connected* if there exists an undirected path from i to j, and we say subsets $A, B \subset V$ are connected if there exists an $i \in A$ and $j \in B$ that are connected.

Definition 2.1.22 (Tree messages)**.** *Given a density with respect to a base measure ν that factorizes on an undirected tree $G = (V, E)$ of the form (2.1.16), we define the* message from node i to node j *with $\{i, j\} \in E$ to be*

$$m_{j\to i}(x_i) \triangleq \int \prod_{\{i',j'\}\in E'} \psi_{i'j'}(x_{i'}, x_{j'})\psi_{ij}(x_i, x_j)\nu(dx_{V'}) \tag{2.1.18}$$

where (V', E') is the subtree of G that is disconnected from node i in $G \setminus \{j\}$.

Theorem 2.1.23 (Tree message passing)**.** *For a density that factorizes on an undirected tree $G = (V, E)$ of the form (2.1.16), the result of any computation of the form (2.1.13) can be written in terms of messages as*

$$\int \prod_{\{i,j\}\in E} \psi_{ij}(x_i, x_j)\nu(dx_A) = \prod_{\{i,j\}\in E'} \psi_{ij}(x_i, x_j) \prod_{k\in B} m_{j\to k}(x_k) \tag{2.1.19}$$

where $(V', E') = G \setminus A$ is the graph over the the nodes that are not integrated out and B is the set of nodes in V' that have edges to A in G, i.e. $B = \{k \in V' : \{k, j\} \in E, j \in A\}$. Furthermore, all messages can be computed efficiently and simultaneously via the recursions

$$m_{i\to j}(x_i) = \int \psi_{ij}(x_i, x_j) \prod_{k\in n_G(j)\setminus\{i\}} m_{j\to k}(x_j)\nu(dx_j). \tag{2.1.20}$$

Therefore all inference queries in undirected trees can be computed in time linear in the length of the longest path in the tree.

Proof. The theorem follows from applying elimination to trees and expressing every partial elimination result as (2.1.20). □

We refer to an implementation of the recursion (2.1.20) as a *tree message-passing algorithm*. We use tree message passing algorithms extensively as subroutines in the inference algorithms for the time series models that we develop in the sequel.

■ 2.2 Exponential families and conjugacy

In this section, we define exponential families and give some of the properties that we use in this thesis.

■ 2.2.1 Definition and basic properties

Definition 2.2.1 (Exponential family)**.** *We say a family of densities p with respect to a base measure ν indexed by a parameter vector θ is an* exponential family *of densities if it can be written as*

$$p(x|\theta) = h(x)\exp\{\langle \eta(\theta), t(x)\rangle - Z(\eta(\theta))\} \tag{2.2.1}$$

where $\langle\cdot,\cdot\rangle$ is an inner product on real vector spaces. We call $\eta(\theta)$ the natural parameter *vector, $t(x)$ the (sufficient)* statistic *vector, $h(x)$ the* base density*, and*

$$Z(\eta) \triangleq \ln \int e^{\langle \eta, t(x)\rangle} h(x)\nu(dx) \tag{2.2.2}$$

the log partition function.

It is often useful to parameterize the family directly in terms of η, in which case we simply write the density as $p(x|\eta)$. Note that marginalizing over part of the support of an exponential family, such as marginalizing over one coordinate of x or of $t(x)$, does not in general yield another exponential family.

Given an exponential family of the form (2.2.1) we define the set of natural parameters that yield valid normalizable probability densities as Θ, where

$$\Theta \triangleq \{\eta : Z(\eta) < \infty\} \tag{2.2.3}$$

and the set of realizable expected statistics as

$$\mathcal{M} \triangleq \{\mathbb{E}_{X\sim p(\cdot|\eta)}[t(X)] : \eta \in \Theta\} \tag{2.2.4}$$

where $X \sim p(\,\cdot\,|\eta)$ denotes that X is distributed with density $p(\,\cdot\,|\eta)$. We say a family is *regular* if Θ is open, and *minimal* if there is no nonzero a such that $\langle a, t(x)\rangle$ is equal

to a constant (ν-a.e.). Minimality ensures that there is a unique natural parameter for each possible density (up to values on sets of ν-measure 0). We say a family is *tractable* if for any η we can evaluate $Z(\eta)$ efficiently[1] and when $X \sim p(\,\cdot\,|\eta)$ we can compute $\mathbb{E}[t(X)]$ and simulate samples of X efficiently.

For a regular exponential family, derivatives of Z are related to expected statistics.

Proposition 2.2.2 (Mean mapping and cumulants)**.** *For a regular exponential family of densities of the form (2.2.1) with $X \sim p(\,\cdot\,|\eta)$, we have $\nabla Z : \Theta \to \mathcal{M}$ and*

$$\nabla Z(\eta) = \mathbb{E}[t(X)] \tag{2.2.5}$$

and writing $\mu \triangleq \mathbb{E}[t(X)]$ we have

$$\nabla^2 Z(\eta) = \mathbb{E}[t(X)t(X)^\mathsf{T}] - \mu\mu^\mathsf{T}. \tag{2.2.6}$$

More generally, the moment generating function of $t(X)$ can be written

$$\mathrm{M}_{t(X)}(s) \triangleq \mathbb{E}[e^{\langle s, t(X)\rangle}] = e^{Z(\eta+s)-Z(\eta)}. \tag{2.2.7}$$

and so derivatives of Z give cumulants *of $t(X)$, where the first cumulant is the mean and the second and third cumulants are the second and third central moments, respectively.*

Proof. Note that we require the family to be regular even to define ∇Z. To show 2.2.5, using ν as the base measure for the density we write

$$\begin{aligned}
\nabla_\eta Z(\eta) &= \nabla_\eta \ln \int e^{\langle \eta, t(x)\rangle} h(x)\nu(dx) && (2.2.8)\\
&= \frac{1}{\int e^{\langle \eta, t(x)\rangle} h(x)\nu(dx)} \int t(x) e^{\langle \eta, t(x)\rangle} h(x)\nu(dx) && (2.2.9)\\
&= \int t(x) p(x|\eta)\nu(dx) && (2.2.10)\\
&= \mathbb{E}[t(X)]. && (2.2.11)
\end{aligned}$$

To derive the form of the moment generating function, we write

[1] We do not provide a precise definition of computational efficiency here. Common definitions often correspond to the complexity classes P or BPP [3].

$$\mathbb{E}[e^{\langle s,t(X)\rangle}] = \int e^{\langle s,t(x)\rangle} p(x)\nu(dx) \tag{2.2.12}$$

$$= \int e^{\langle s,t(x)\rangle} e^{\langle \eta,t(x)\rangle - Z(\eta)} h(x)\nu(dx) \tag{2.2.13}$$

$$= e^{Z(\eta+s)-Z(\eta)}. \tag{2.2.14}$$

The cumulant generating function for $t(X)$ is then $\ln \mathrm{M}_{t(X)}(s) = Z(\eta + s) - Z(\eta)$. □

When a specific set of expected statistics can only arise from one member of an exponential family, we say the family is *identifiable* and we can use the moments as an alternative way to parameterize the family.

Theorem 2.2.3 (Exponential family identifiability)**.** *For a regular, minimal exponential family of the form (2.2.1), $\nabla Z : \Theta \to \mathcal{M}$ is injective, and $\nabla Z : \Theta \to \mathcal{M}^\circ$ is surjective, where $\mathcal{M}^\circ$ denotes the interior of $\mathcal{M}$. Therefore $\nabla Z : \Theta \to \mathcal{M}^\circ$ is a bijection.*

Proof. See Wainwright and Jordan [113, Theorem 3.3]. □

When parameterizing a regular, minimal exponential family in terms of expected statistics $\mu \in \mathcal{M}^\circ$, we say it is written with *mean parameters*, and we have $\eta(\mu) = (\nabla Z)^{-1}(\mu)$ using Theorem 2.2.3. Given a set of moments, the corresponding minimal exponential family member has a natural interpretation as the density (relative to ν) with maximum entropy subject to those moment constraints [113, Section 3.1].

For members of an exponential family, many quantities can be expressed generically in terms of the natural parameter, expected statistics under that parameter, and the log partition function.

Proposition 2.2.4 (Entropy, score, and Fisher information)**.** *For a regular exponential family of densities of the form (2.2.1) parameterized in terms of its natural parameter η, with $X \sim p(\,\cdot\,|\eta)$ and $\mu(\eta) \triangleq \mathbb{E}[t(X)]$ we have*

1. *The (differential)* entropy *is*

$$\mathbb{H}[p] \triangleq -\mathbb{E}[\ln p(X|\eta)] = -\langle \eta, \mu(\eta)\rangle + Z(\eta). \tag{2.2.15}$$

2. *When the family is regular, the* score *with respect to the natural parameter is*

$$v(x,\eta) \triangleq \nabla_\eta \ln p(x|\eta) = t(x) - \mu(\eta) \tag{2.2.16}$$

3. *When the family is regular, the* Fisher information *with respect to the natural parameter is*

$$\mathcal{I}(\eta) \triangleq \mathbb{E}[v(X,\eta)v(X,\eta)^{\mathsf{T}}] = \nabla^2 Z(\eta). \tag{2.2.17}$$

Proof. Each follows from (2.2.1), where (2.2.16) and (2.2.17) use Proposition 2.2.2. □

When the family is parameterized in terms of some other parameter θ so that $\eta = \eta(\theta)$, the properties in Proposition 2.2.4 include Jacobian terms of the form $\partial\eta/\partial\theta$. When the alternative parameter is the mean parameter μ, since $\eta(\mu) = (\nabla Z)^{-1}(\mu)$ the relevant Jacobian is $(\nabla^2 Z)^{-1}$. We use the Fisher information expression (2.2.17) in the development of a stochastic variational inference (SVI) algorithm in Chapter 5.

We conclude this subsection with two examples of exponential families of densities.

Example 2.2.5 (Gaussian)**.** *The Gaussian PDF can be parameterized in terms of its mean and covariance (μ, Σ), where $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ with $\Sigma \succ 0$, and can be written*

$$\begin{aligned} p(x|\mu,\Sigma) &= (2\pi)^{-d/2}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(x-\mu)^{\mathsf{T}}\Sigma^{-1}(x-\mu)\right\} \\ &= \underbrace{(2\pi)^{-d/2}}_{h_{\mathrm{N}}(x)}\exp\Big\{\Big\langle \underbrace{(-\tfrac{1}{2}\Sigma^{-1}, \Sigma^{-1}\mu)}_{\eta_{\mathrm{N}}(\mu,\Sigma)}, \underbrace{(xx^{\mathsf{T}}, x)}_{t_{\mathrm{N}}(x)}\Big\rangle - \underbrace{(\tfrac{1}{2}\ln|\Sigma| + \tfrac{1}{2}\mu^{\mathsf{T}}\Sigma^{-1}\mu)}_{Z_{\mathrm{N}}(\eta_{\mathrm{N}}(\mu,\Sigma))}\Big\}. \end{aligned}$$

where $\langle (A,b),(C,d)\rangle = \mathrm{tr}(A^{\mathsf{T}}C) + b^{\mathsf{T}}d$. Therefore it is an exponential family of densities, and further it is a regular exponential family since $\Theta = \{(\Sigma,\mu) : \Sigma \succ 0, \mu \in R^d\}$ is open (in the standard product topology for Euclidean spaces). The family is minimal because there is no nonzero (A,b) such that $x^{\mathsf{T}}Ax + b^{\mathsf{T}}x$ equal to a constant (ν-a.e.).

Example 2.2.6 (Categorical)**.** *Consider drawing a sample X from a finite distribution with support on $[K] = \{1,2,\ldots,K\}$, where $p(x = k|\pi) = \pi_k$ for $K \in [K]$ and π satisfies $\sum_i \pi_i = 1$ and $\pi > 0$ element-wise. We can write the PMF for X as*

$$p(x|\bar{\pi}) = \prod_{k=1}^{K} \pi_k^{\mathbb{I}[x=k]} = \exp\Big\{\sum_{k=1}^{K} \ln\pi_k \mathbb{I}[x=k]\Big\} = \exp\big\{\langle \ln\pi, \mathbb{1}_x\rangle\big\} \tag{2.2.18}$$

where the log is taken element-wise, $\mathbb{I}[\,\cdot\,]$ is an indicator function that takes value 1 when its argument is true and 0 otherwise, and $\mathbb{1}_k$ is an indicator vector with its ith entry $\mathbb{I}[k=i]$. We call this family of densities the categorical *family, and it is an exponential family of densities with natural parameter $\eta(\pi) = \ln\pi$ and statistic $t(x) = \mathbb{1}_x$. (A closely related family is the* multinomial *family, where we consider drawing a set of n independent samples from the same process, $x = \{x_i : i \in [n]\}$, and defining the statistic to be the counts of each occurrence, i.e. the kth entry of $t(x)$ is $|\{x_i : x_i = k\}|$.)*

Note that $Z(\pi) = 0$. The categorical family as written in (2.2.18) is not a regular exponential family because $\Theta = \{\pi \in \mathbb{R}^K : \sum_i \pi_i = 1,\ \pi > 0\}$ is not open. Since the family is not regular, (2.2.16) does not apply. We can instead write the family of densities as

$$p(x|\pi) = \exp\left\{\langle \ln \bar{\pi}, \mathbb{1}_x \rangle - \ln \textstyle\sum_{i=1}^K \bar{\pi}_i\right\} \tag{2.2.19}$$

where $Z(\eta(\bar{\pi})) = \ln \sum_{i=1}^K \bar{\pi}_i$ so that $\Theta = \{\bar{\pi} \in \mathbb{R}^K : \pi > 0\}$ is open. However, neither (2.2.18) or (2.2.19) is a minimal exponential family, since the statistic satisfies $\langle 1, \mathbb{1}_x \rangle = 1$ for any $x \in [K]$. As is clear from the renormalization in (2.2.19), the parameter is not identifiable, so Theorem 2.2.3 does not apply.

While it is possible to write the categorical as a regular minimal exponential family by removing one component from the parameterization, it is often easiest to work with the categorical (and multinomial) in the form (2.2.18).

■ 2.2.2 Conjugacy

In this subsection, we define a notion of conjugacy for pairs of families of distributions. Conjugate families are especially useful for Bayesian analysis and algorithms.

Definition 2.2.7. *Given two (not necessarily exponential) families of densities $p_1(\theta|\alpha)$ and $p_2(x|\theta)$ indexed by parameters α and θ, respectively, we say the pair (p_1, p_2) are a* conjugate pair of densities *if for all α, x, and θ we have*

$$p_1(\theta|\alpha)p_2(x|\theta) \propto p_1(\theta|\alpha') \tag{2.2.20}$$

for some $\alpha' = \alpha'(x, \alpha)$ that may depend on x and α.

We can extend this definition to distributions without densities by considering instead indexed families of laws [88, Definition 2].

Conjugate pairs are particularly useful in Bayesian analysis because if we have a *prior family* $p(\theta|\alpha)$ and we observe data generated according to a *likelihood* $p(x|\theta)$ then the *posterior* $p(\theta|x, \alpha)$ is in the same family as the prior. In the context of Bayesian updating, we call α the *hyperparameter* and α' the *posterior hyperparameter*.

Given a regular exponential family likelihood, we can always define a conjugate prior, as shown in the next proposition.

Proposition 2.2.8. *Given a regular exponential family*

$$\begin{aligned} p_{X|\theta}(x|\theta) &= h_X(x) \exp\{\langle \eta_X(\theta), t_X(x) \rangle - Z_X(\eta_X(\theta)) && (2.2.21) \\ &= h_X(x) \exp\{\langle (\eta_X(\theta), -Z_X(\eta(\theta))),\ (t_X(x), 1) \rangle\} && (2.2.22) \end{aligned}$$

then if we define the statistic $t_\theta(\theta) \triangleq (\eta_X(\theta), -Z_X(\eta(\theta)))$ *and an exponential family of densities with respect to that statistic as*

$$p_{\theta|\alpha}(\theta|\alpha) = h_\theta(\theta) \exp\{\langle \eta_\theta(\alpha), t_\theta(\theta)\rangle - Z_\theta(\eta_\theta(\alpha))\} \tag{2.2.23}$$

then the pair $(p_{\theta|\alpha}, p_{X|\theta})$ *is a conjugate pair of families with*

$$p(\theta|\alpha)p(x|\theta) \propto h_\theta(\theta) \exp\{\langle \eta_\theta(\alpha) + (t_X(x), 1), t_\theta(\theta)\rangle\} \tag{2.2.24}$$

and hence we can write the posterior hyperparameter as $\alpha' = \eta_\theta^{-1}(\eta_\theta(\alpha) + (t_X(x), 1))$. *When the prior family is parameterized with its natural parameter, we have* $\eta' = \eta + (t_X(x), 1)$.

As a consequence of Proposition 2.2.8, if the prior family is written with natural parameters and we generate data $\{x_i\}_{i=1}^n$ according to the model

$$\theta \sim p_{\theta|\eta}(\,\cdot\,|\eta) \tag{2.2.25}$$

$$x_i|\theta \overset{\text{iid}}{\sim} p_{X|\theta}(\,\cdot\,|\theta) \quad i = 1, 2, \ldots, n, \tag{2.2.26}$$

where the notation $x_i \overset{\text{iid}}{\sim} p(\,\cdot\,)$ denotes that the random variables x_i are independently and identically distributed, then $p(\theta|\{x_i\}_{i=1}^n, \eta)$ has posterior hyperparameter $\eta' = \eta + (\sum_{i=1}^n t(x_i), n)$. Therefore if a prior family $p(\,\cdot\,|\eta)$ is tractable then the posterior under the conjugate likelihood is tractable.

We conclude this subsection with two examples of conjugate pairs of exponential families. A list of conjugate pairs can be found in Gelman et al. [38].

Example 2.2.9 (NIW-Gaussian conjugacy)**.** *Here we show that the normal-inverse-Wishart (NIW) is a conjugate prior family for the Gaussian likelihood of Example 2.2.5.*

The NIW density with parameter $\alpha = (\Lambda_0, \mu_0, \kappa_0, \nu_0)$ *and* $\Lambda_0 \succ 0$, $\kappa_0 > 0$, $\nu_0 > d$ *is*

$$\begin{aligned} p(\mu, \Sigma|\alpha) &\propto |\Sigma|^{-((\nu_0+d)/2+1)} \exp\Big\{ -\frac{1}{2}\operatorname{tr}(\Lambda_0\Sigma^{-1}) - \frac{\kappa_0}{2}(\mu-\mu_0)^\mathsf{T}\Sigma^{-1}(\mu-\mu_0)\Big\} \\ &= \underbrace{|\Sigma|^{-d/2+1}}_{h_{\text{NIW}}(\mu,\Sigma)} \exp\Big\{\Big\langle \underbrace{(\Lambda_0 + \kappa_0\mu_0\mu_0^\mathsf{T}, \kappa_0\mu_0, \kappa_0, \nu_0)}_{\eta_{\text{NIW}}(\Lambda_0,\mu_0,\kappa_0,\nu_0)}, \underbrace{(-\tfrac{1}{2}\Sigma^{-1}, \Sigma^{-1}\mu, -\tfrac{1}{2}\mu^\mathsf{T}\Sigma^{-1}\mu, -\tfrac{1}{2}\ln|\Sigma|)}_{t_{\text{NIW}}(\mu,\Sigma)} \Big\rangle\Big\} \end{aligned}$$

where $\langle (A, b, c, d), (E, f, g, h)\rangle = \operatorname{tr}(A^\mathsf{T}E) + b^\mathsf{T}f + cg + dh$.

The normal density from Example 2.2.5 is

$$\begin{aligned} p(x|\mu,\Sigma) &= \underbrace{(2\pi)^{-d/2}}_{h_{\mathrm{N}}(x)} \exp\left\{\left\langle \underbrace{(-\tfrac{1}{2}\Sigma^{-1}, \Sigma^{-1}\mu)}_{\eta_{\mathrm{N}}(\mu,\Sigma)}, \underbrace{(xx^{\mathsf{T}}, x)}_{t_{\mathrm{N}}(x)} \right\rangle - \underbrace{(\tfrac{1}{2}\ln|\Sigma| + \tfrac{1}{2}\mu^{\mathsf{T}}\Sigma^{-1}\mu)}_{Z_{\mathrm{N}}(\eta_{\mathrm{N}}(\mu,\Sigma))} \right\} \\ &= \underbrace{(2\pi)^{-d/2}}_{h_{\mathrm{N}}(x)} \exp\left\{\left\langle \underbrace{(-\tfrac{1}{2}\Sigma^{-1}, \Sigma^{-1}\mu, -\tfrac{1}{2}\mu^{\mathsf{T}}\Sigma^{-1}\mu, -\tfrac{1}{2}\ln|\Sigma|)}_{(\eta_{\mathrm{N}}(\mu,\Sigma), -Z_{\mathrm{N}}(\eta_{\mathrm{N}}(\mu,\Sigma)))}, \underbrace{(xx^{\mathsf{T}}, x, 1, 1)}_{(t_{\mathrm{N}}(x),1)} \right\rangle\right\}. \end{aligned}$$

The joint likelihood of n independent samples $\{x_i \overset{iid}{\sim} \mathcal{N}(\mu,\Sigma) : i \in [n]\}$ is

$$\begin{aligned} p(\{x\}_{i=1}^n|\mu,\Sigma) &= \prod_{i=1}^n p(x_i|\mu,\Sigma) \\ &= (2\pi)^{-nd/2} \exp\left\{ -\frac{1}{2}\left\langle \left(\eta_{\mathrm{N}}(\mu,\Sigma), -Z_{\mathrm{N}}(\eta_{\mathrm{N}}(\mu,\Sigma))\right), \textstyle\sum_{i=1}^n \left(t_{\mathrm{N}}(x_i), 1\right) \right\rangle\right\} \end{aligned}$$

so we see that $p(\mu,\Sigma|\{x\}_{i=1}^n,\alpha)$ is in the NIW family with a new natural parameter

$$\begin{aligned} \eta_{\mathrm{NIW}}(\Lambda_n,\mu_n,\kappa_n,\nu_n) &= \eta_{\mathrm{NIW}}(\Lambda_0,\mu_0,\kappa_0,\nu_0) + \textstyle\sum_{i=1}^n \left(t_{\mathrm{N}}(x_i),1\right) \\ &= (\Lambda_0 + \kappa_0\mu_0\mu_0^{\mathsf{T}}, \kappa_0\mu_0, \kappa_0, \nu_0) + (\textstyle\sum_i x_i x_i^{\mathsf{T}}, \sum_i x_i, n, n) \end{aligned}$$

where it can be checked by writing η_{NIW}^{-1} that

$$\begin{aligned} \kappa_n &= \kappa_0 + n \\ \nu_n &= \nu_0 + n \\ \mu_n &= \frac{\kappa_0}{\kappa_0+n} + \frac{n}{\kappa_0+n}\bar{x} \\ \Lambda_n &= \Lambda_0 + S + \frac{\kappa_0 n}{\kappa_0+n}(\bar{x}-\mu_0)(\bar{x}-\mu_0)^{\mathsf{T}} \end{aligned}$$

with $\bar{x} = \frac{1}{n}\sum_i x_i$ and $S = \sum_i (x_i-\bar{x})(x_i-\bar{x})^{\mathsf{T}}$ as in Gelman et al. [38].

Example 2.2.10 (Dirichlet-categorical conjugacy)**.** *Here we show that the Dirichlet is a conjugate prior family for the categorical likelihood of Example 2.2.6.*

The K-dimensional Dirichlet density with parameter $\alpha \in \mathbb{R}_+^K$ where $\alpha > 0$ can be written

$$p(\pi|\alpha) = \mathrm{Dir}(\alpha) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K x_i^{\alpha_i - 1} \tag{2.2.27}$$

$$\propto \exp\{\langle \alpha - 1, \ln \pi \rangle\}. \tag{2.2.28}$$

Using the categorical density we have

$$p(\pi|x,\alpha) \propto p(\pi|\alpha)p(x|\pi) \tag{2.2.29}$$
$$\propto \exp\{\langle \alpha - 1, \ln \pi \rangle\} \exp\{\langle \ln \pi, \mathbb{1}_x \rangle\} \tag{2.2.30}$$
$$\propto \exp\{\langle \alpha - 1 + \mathbb{1}_x, \ln \pi \rangle\} \tag{2.2.31}$$

where, as in Example 2.2.6, $\mathbb{1}_x$ is an indicator vector with its xth entry set to 1 and its other entries 0. Therefore the posterior $p(\pi|x,\alpha)$ is in the Dirichlet family with parameter $\alpha + \mathbb{1}_x$. Similarly, for the Multinomial likelihood

$$p(x|\pi) = \exp\{\langle \ln \pi, n_x \rangle\} \tag{2.2.32}$$

where $n_x = \sum_i \mathbb{1}_{x_i}$ so that the jth component is the number of occurences of outcome j, the posterior $p(\pi|x,\alpha)$ is in the Dirichlet family with parameter $\alpha + n_x$.

■ 2.3 Bayesian inference algorithms in graphical models

Here we outline the standard Bayesian inference algorithms and how they relate to the graphical model structure described in Section 2.1 as well as the exponential family conjugacy structure described in Section 2.2. In particular, we describe algorithms that are *compositional* in terms of graphical model structure and that have particularly efficient updates when the graphical model is itself composed of tractable exponential family distributions.

■ 2.3.1 Gibbs sampling

In Gibbs sampling, and sampling methods more generally, the task is to generate samples from a distribution of interest so that any probability or statistic can be estimated using the sample population. For a Markov Chain Monte Carlo (MCMC) method such as Gibbs sampling, to generate samples for some collection of random variables X the algorithm simulates a Markov chain on the range of X such that the *limiting distribution* or *stationary distribution* of the chain is the target distribution of X. In the Bayesian context, the distribution of interest is typically an intractable posterior.

Given a collection of n random variables $X = \{X_i : i \in [n]\}$, the Gibbs sampling algorithm iteratively samples each variable conditioned on the sampled values of the others. When the random variables are Markov on a graph $G = (V, E)$, the conditioning can be reduced to each variable's respective Markov blanket, as in Algorithm 2.1.

A variant of the *systematic scan* of Algorithm 2.1, in which nodes are traversed in a fixed order for each outer iteration, is the *random scan*, in which nodes are traversed according to a random permutation sampled for each outer iteration. An advantage of the random scan (and other variants) is that the chain becomes reversible and therefore

Algorithm 2.1 Gibbs sampling

Input: distribution X on graph G with N nodes, conditionals $p_{X_i|X_{\mathrm{MB}_G(i)}}(x_i|x_{\mathrm{MB}_G(i)})$
Output: samples $\{\hat{x}^{(t)}\}$
  Initialize $x = (x_1, x_2, \ldots, x_N)$
  for $t = 1, 2, \ldots$ **do**
    for $i = 1, 2, \ldots, N$ **do**
      $x_i \sim p_{X_i|X_{\mathrm{MB}_G(i)}}(\,\cdot\,|x_{\mathrm{MB}_G(i)})$
    $\hat{x}^{(t)} \leftarrow (x_1, x_2, \ldots, x_N)$

simpler to analyze [94, Section 10.1.2]. With the conditional independencies implied by a graph, some sampling steps may be performed in parallel.

A Markov chain is called *ergodic* if has a unique steady-state distribution for any initial state; see Robert and Casella [94, Section 6.6.1] for a definition for countable state spaces and Meyn and Tweedie [76, Chapter 13] for a definition for general state spaces. If the Markov chain produced by a Gibbs sampling algorithm is ergodic, then the stationary distribution is the target distribution of X [94, Theorem 10.6]. The Markov chain for a Gibbs sampler can fail to be ergodic if, for example, the support of the target distribution is disconnected [94, Example 10.7]. Many of the Gibbs samplers we develop result in ergodic chains because all of the conditional densities exist and are positive [94, Theorem 10.8]. The main performance criterion of an MCMC sampler is its *mixing time* [94, Chapter 12], which measures the rate at which the distribution of the chain's state reaches the target distribution.

For a more detailed treatment of Gibbs sampling theory, see Robert and Casella [94, Chapter 6, Chapter 10]. For a detailed treatment of Markov chain ergodic theory for general state spaces, as required in precise treatments of Gibbs samplers for the Dirichlet process, see Meyn and Tweedie [76].

■ 2.3.2 Mean field variational inference

In mean field, and variational inference more generally, the task is to approximate an intractable distribution, such as a complex posterior, with a distribution from a tractable family in which inference can be performed efficiently. In this section we define the mean field optimization problem and derive the standard coordinate optimization algorithm. We also give some basic results on the relationship between mean field and both graphical model and exponential family structure. For concreteness and simpler notation, we work mostly with undirected graphical models; the results extend immediately to directed models.

Mean field inference makes use of several densities and distributions, and so we use a subscript notation for expectations to clarify the measure used in the integration when

it cannot easily be inferred from context. Given a function f and a random variable X with range $\mathcal{X}$ and density p with respect to a base measure ν, we write the expectation of f as

$$\mathbb{E}_{p(X)}\left[f(X)\right] = \int_{\mathcal{X}} f(x)p(x)\nu(dx). \tag{2.3.1}$$

Proposition 2.3.1 (Variational inequality)**.** *For a density p with respect to a base measure ν of the form*

$$p(x) = \frac{1}{Z}\bar{p}(x) \quad \textit{with} \quad Z \triangleq \int \bar{p}(x)\nu(dx), \tag{2.3.2}$$

for all densities q with respect to ν we have

$$\ln Z = \mathcal{L}[q] + \mathrm{KL}(q||p) \geq \mathcal{L}[q] \tag{2.3.3}$$

where

$$\mathcal{L}[q] \triangleq \mathbb{E}_{q(X)}\left[\ln \frac{\bar{p}(X)}{q(X)}\right] = \mathbb{E}_{q(X)}\left[\ln \bar{p}(X)\right] + \mathbb{H}[q] \tag{2.3.4}$$

$$\mathrm{KL}(q||p) \triangleq \mathbb{E}_{q(X)}\left[\ln \frac{q(X)}{p(X)}\right]. \tag{2.3.5}$$

Proof. To show the equality, with $X \sim q$ we write

$$\mathcal{L}[q] + \mathrm{KL}(q||p) = \mathbb{E}_{q(X)}\left[\frac{\bar{p}(X)}{q(X)}\right] + \mathbb{E}_{q(X)}\left[\ln \frac{q(X)}{p(X)}\right] \tag{2.3.6}$$

$$= \mathbb{E}_{q(X)}\left[\ln \frac{\bar{p}(X)}{p(X)}\right] \tag{2.3.7}$$

$$= \ln Z. \tag{2.3.8}$$

The inequality follows from the property $\mathrm{KL}(q||p) \geq 0$, known as Gibbs's inequality, which follows from Jensen's inequality and the fact that the logarithm is concave:

$$-\mathrm{KL}(q||p) = \mathbb{E}_{q(X)}\left[\ln \frac{p(X)}{q(X)}\right] \leq \ln \int q(x)\frac{p(x)}{q(x)}\nu(dx) = 0 \tag{2.3.9}$$

with equality if and only if $q = p$ (ν-a.e.). □

We call the log of $\bar{p}$ in (2.3.2) the *energy* and $\mathcal{L}[q]$ the *variational lower bound*, and say $\mathcal{L}[q]$ decomposes into the average energy plus entropy as in (2.3.4). For two densities q and p with respect to the same base measure, $\mathrm{KL}(q||p)$ is the *Kullback-Leibler*

divergence from q to p, used as a measure of dissimilarity between pairs of densities [2].

The variational inequality given in Proposition 2.3.1 is useful in inference because if we wish to approximate an intractable p with a tractable q by minimizing $\mathrm{KL}(q||p)$, we can equivalently choose q to maximize $\mathcal{L}[q]$, which is possible to evaluate since it does not include the partition function Z.

In the context of Bayesian inference, p is usually an intractable posterior distribution of the form $p(\theta|x,\alpha)$, $\bar{p}$ is the unnormalized product of the prior and likelihood $\bar{p}(\theta) = p(\theta|\alpha)p(x|\theta)$, and Z is the marginal likelihood $p(x|\alpha) = \int p(x|\theta)p(\theta|\alpha)\nu(d\theta)$, which plays a central role in Bayesian model selection and the minimum description length (MDL) criterion [74, Chapter 28] [49, Chapter 7].

Given that graphical model structure can affect the complexity of probabilistic inference, as discussed in Section 2.1.3, it is natural to consider families q that factor according to tractable graphs.

Definition 2.3.2 (Mean field variational inference)**.** *Let p be the density with respect to ν for a collection of random variables $X = (X_i : i \in V)$, and let*

$$\mathcal{Q} \triangleq \{q : q(x) \propto \prod_{C \in \mathcal{C}} q_C(x_C)\} \tag{2.3.10}$$

be a family of densities with respect to ν that factorize according to a graph $G = (V, E)$ with $\mathcal{C}$ the set of maximal cliques of G. Then the mean field optimization problem *is*

$$q^* = \arg\max_{q \in \mathcal{Q}} \mathcal{L}[q] \tag{2.3.11}$$

where $\mathcal{L}[q]$ is defined as in (2.3.4).

Note that the optimization problem is not convex and so one can only expect to find a local optimum of the objective [113]. However, since the objective is convex in each q_C individually, an optimization procedure that updates each factor in turn holding the rest constant will converge to a local optimum [11, Chapter 10]. We call such a coordinate ascent procedure on (2.3.11) a *mean field algorithm*. For approximating families in a factored form, we can derive a generic update to be used in a mean field algorithm.

Proposition 2.3.3 (Mean field update)**.** *Given a mean field objective as in Definition 2.3.2, the optimal update to a factor q_A fixing the other factors defined by $q_A^* = \arg\max_{q_A} \mathcal{L}[q]$ is*

$$q_A^*(x_A) \propto \exp\{\mathbb{E}[\ln \bar{p}(x_A, X_{A^c})]\} \tag{2.3.12}$$

where the expectation is over $X_{A^c} \sim q_{A^c}$ with $q_{A^c}(x_{A^c}) \propto \prod_{C \in \mathcal{C} \setminus A} q_C(x_C)$.

Proof. Dropping terms constant with respect to q_A, we write

$$q_A^* = \arg\min_{q_A} \mathrm{KL}(q||p) \tag{2.3.13}$$

$$= \arg\min_{q_A} \mathbb{E}_{q_A}\left[\ln q_A(X_A)\right] + \mathbb{E}_{q_A}\left[\mathbb{E}_{q_{A^c}}\left[\log \bar{p}(X)\right]\right] \tag{2.3.14}$$

$$= \arg\min_{q_A} \mathrm{KL}(q_A||\widetilde{p}_A) \tag{2.3.15}$$

where $\widetilde{p}_A(x_A) \propto \exp\{\mathbb{E}_{q_{A^c}}[\ln \bar{p}(x_A, X_{A^c})]\}$. Therefore, we achieve the unique (ν-a.e.) minimum by setting $q_A = \widetilde{p}_A$. □

Furthermore, if p factors according to a graph then the same graph structure is induced in the factors of the optimal q.

Proposition 2.3.4 (Induced graph structure)**.** *If p is Markov on a graph $G = (V, E)$ with the form $p(x) \propto \prod_{C\in\mathcal{C}} \psi_C(x_C)$ for cliques $\mathcal{C}$ of G, then any optimal factor q_A with $A \subseteq V$ factors according to $G \setminus A^c$. Furthermore, the update (2.3.12) can be computed using only the factors on the cliques $\mathcal{C}' = \{C \in \mathcal{C} : C \cap A \neq \varnothing\}$, i.e. the cliques on variables in the Markov blanket of A.*

Proof. Using (2.3.12) we have

$$q_A^*(x_A) \propto \exp\{\mathbb{E}[\ln \bar{p}(x_A, X_{A^c})]\} \propto \exp\left\{\sum_{C\in\mathcal{C}'} \mathbb{E}[\ln \psi_C(X_C)]\right\} \tag{2.3.16}$$

where factors not involving the variables in A are dropped up to proportionality. □

Because q inherits the graphical structure of p, it is therefore natural to consider tractable families $\mathcal{Q}$ that are Markov on *subgraphs* of p. Note that when q is a subgraph of p, the variational lower bound is a sum of terms corresponding to the factors of p. When the family $\mathcal{Q}$ is chosen to be Markov on the completely disconnected graph $G = (V, E)$ with $E = \varnothing$, the resulting algorithm is called *naive mean field*. When the tractable subgraph retains some nontrivial graphical structure, the algorithm is called *structured mean field*. In this thesis we use structured mean field extensively for inference in time series models.

Finally, we note the simple form of updates for exponential family conjugate pairs.

Proposition 2.3.5 (Mean field and conjugacy)**.** *If x_i appears in $\bar{p}$ only in an exponential family conjugate pair (p_1, p_2) where*

$$p_1(x_i|x_{\pi_G(i)}) \propto \exp\{\langle \eta(x_{\pi_G(i)}), t(x_i)\rangle\} \tag{2.3.17}$$
$$p_2(x_{c_G(i)}|x_i) = \exp\{\langle t(x_i), (t(x_{c_G(i)}), 1)\rangle\} \tag{2.3.18}$$

then the optimal factor $q_i(x_i)$ is in the prior family with natural parameter

$$\widetilde{\eta} \triangleq \mathbb{E}_q[\eta(X_{\pi_G(i)})] + \mathbb{E}_q[(t(X_{c_G(i)}), 1)]. \tag{2.3.19}$$

Proof. The result follows from substituting (2.3.17) and (2.3.18) into (2.3.12). □

See Wainwright and Jordan [113, Chapter 5] for a convex analysis perspective on mean field algorithms in graphical models composed of exponential families.

■ 2.4 Hidden Markov Models

In this section we use the definitions and results from previous sections to define the Bayesian Hidden Markov Model and give Gibbs sampling and mean field inference algorithms. For simplicity, we refer only to a single observation sequence; the generalization to multiple observation sequences is immediate.

A Hidden Markov Model (HMM) on N states defines a joint distribution over a *state sequence* $x_{1:T}$ and an *observation sequence* $y_{1:T}$. It is parameterized by an *initial state distribution* $\pi^{(0)} \in \mathbb{R}_+^N$, a *transition matrix* $A \in \mathbb{R}_+^{N\times N}$, and emission parameters $\theta = \{\theta^{(i)}\}_{i=1}^N$. We use $\pi^{(i)}$ to denote the ith row of A and collect the transition rows and initial state distribution into $\pi = \{\pi^{(i)}\}_{i=0}^N$ for convenient notation.

Definition 2.4.1. *We say sequences of random variables $(x_{1:T}, y_{1:T})$ are distributed according to a Hidden Markov Model, and write $(x_{1:T}, y_{1:T}) \sim \mathrm{HMM}(\pi, \theta)$, when they follow the generative process*

$$x_1 \sim \pi^{(0)}, \tag{2.4.1}$$
$$x_{t+1}|x_t \sim \pi^{(x_t)} \qquad t = 1, 2, \ldots, T-1, \tag{2.4.2}$$
$$y_t|x_t \sim p(\,\cdot\,|\theta^{(x_t)}) \qquad t = 1, 2, \ldots, T. \tag{2.4.3}$$

Figure 2.3 shows a graphical model for the HMM.

If each $p(y|\theta^{(i)})$ is an exponential family of densities in natural parameters of the form

$$p(y|\eta^{(i)}) = \exp\{\langle \eta_y^{(i)}, t_y^{(i)}(y)\rangle - Z_y(\eta_y^{(i)})\} \tag{2.4.4}$$

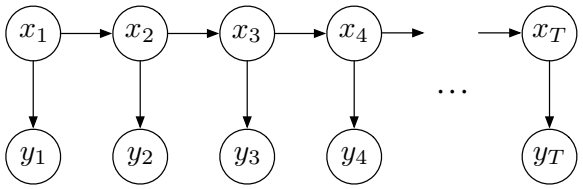

Figure 2.3: Directed graphical model for an HMM.

then we can write the joint density as an exponential family:

$$p(x_{1:T}, y_{1:T}) = \exp\left\{ (\ln \pi^{(0)})^\mathsf{T} \mathbb{1}_{x_1} + \sum_{t=1}^{T-1} \mathbb{1}_{x_t}^\mathsf{T} A \mathbb{1}_{x_{t+1}} + \sum_{t=1}^{T} \langle \eta_y^{(i)}, \mathbb{I}[x_t = i] \cdot t_y^{(i)}(y_t) \rangle \right\}. \tag{2.4.5}$$

Since the HMM is Markov on an undirected tree (more precisely, a chain), we can use the tree message-passing algorithm to perform inference efficiently. The HMM messages and recursions are typically written in terms of *forward messages* F and *backward messages* B, where

$$F_{t,i} \triangleq p(y_{1:t}, x_t) = \sum_{j=1}^{N} A_{ji} F_{t-1,j} p(y_t | \theta^{(i)}) \tag{2.4.6}$$

$$B_{t,i} \triangleq p(y_{t+1:T} | x_t = i) = \sum_{j=1}^{N} A_{ij} p(y_{t+1} | \theta^{(j)}) B_{t+1,j} \tag{2.4.7}$$

with the initial values $F_{1,i} = \pi_i^{(0)} p(y_1 | \theta^{(i)})$ and $B_{T,i} = 1$. Algorithms to compute these messages are given in Algorithms 2.2 and 2.3.

A Bayesian treatment of HMMs places priors on the parameters (π, θ) and includes them in the probabilistic model. We use $\mathrm{Dir}(\alpha)$ where $\alpha \in \mathbb{R}_+^K$ for some $K > 0$ to denote the Dirichlet distribution with parameter α.

Definition 2.4.2. *We say $(\theta, \pi, x_{1:T}, y_{1:T})$ are distributed according to a Bayesian Hidden Markov Model with hyperparameters $\alpha = \{\alpha^{(i)}\}_{i=1}^N$ and $\lambda = \{\lambda^{(i)}\}_{i=1}^N$, and write $(\theta, \pi, x_{1:T}, y_{1:T}) \sim \mathrm{BayesHMM}(\alpha, \lambda)$, when they follow the generative process*

$$\pi^{(i)} \overset{iid}{\sim} \mathrm{Dir}(\alpha^{(i)}) \tag{2.4.8}$$

$$\theta^{(i)} \overset{iid}{\sim} p(\,\cdot\,|\lambda^{(i)}) \tag{2.4.9}$$

$$(x_{1:T}, y_{1:T}) | \pi, \theta \sim \mathrm{HMM}(\pi, \theta) \tag{2.4.10}$$

where $\mathrm{HMM}(\pi, \theta)$ is defined in Definition 2.4.1. Figure 2.4 shows a graphical model.

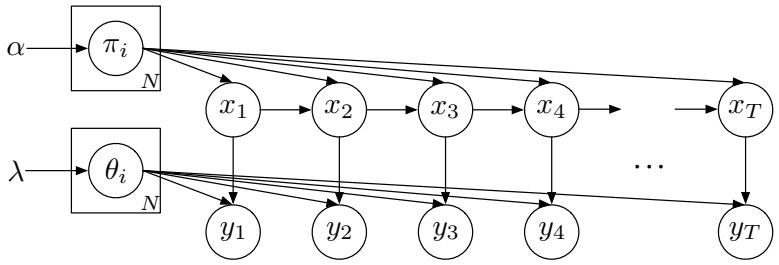

Figure 2.4: Directed graphical model for a Bayesian HMM.

Given an observation sequence $\bar{y}_{1:T}$ the task of interest is to perform inference in the posterior $\pi, \theta, x_{1:T} | \bar{y}_{1:T}$. In this section we develop both Gibbs sampling and mean field algorithms for this inference task.

■ 2.4.1 HMM Gibbs sampling

The HMM Gibbs sampler iterates sampling θ, π, and $x_{1:T}$ from their respective conditionals. To sample the state sequence $x_{1:T}$ from its conditional, we exploit the tree message-passing algorithm. Furthermore, since the Dirichlet is the conjugate prior to the categorical from which each state is sampled, the conditional for π is also Dirichlet. The HMM Gibbs sampler can then be written as Algorithm 2.4.

An alternative Gibbs sampler can be constructed by marginalizing the parameters (π, θ), which is tractable when the observation prior and likelihood form a conjugate pair, and generating samples of $x_{1:T} | \bar{y}_{1:T}, \alpha, \lambda$. While such collapsed samplers can be advantageous in some settings, in the case of a Bayesian HMM eliminating π and θ induces a full graph on the remaining nodes, and so one cannot exploit tree message passing to construct a joint sample of the state sequence and each x_t must be resampled one at a time. Because the x_t are highly correlated in the model, the collapsed Gibbs sampler is often slow to explore the posterior.

■ 2.4.2 HMM mean field

Here we briefly overview a mean field algorithm for HMMs. For more details on the HMM mean field algorithm, see Beal [6, Chapter 3].

We choose a variational family that factorizes as $q(\pi, \theta, x_{1:T}) = q(\pi, \theta) q(x_{1:T})$ so that the parameters and state sequence are decoupled. The Bayesian HMM graphical model then induces independences so that the variational family is

$$q(\pi, \theta, x_{1:T}) = \prod_{i=0}^{N} q(\pi^{(i)}) q(\theta^{(i)}) q(x_{1:T}). \tag{2.4.11}$$

Note that because $\pi^{(i)}$ is the ith row of the transition matrix A, $i = 1, 2, \ldots, N$, we write $q(\pi^{(1)}, \ldots, \pi^{(N)})$ equivalently as $q(A)$ to simplify some notation. The variational objective function is

$$\mathbb{E}\left[\ln \frac{p(\pi, \theta, x, y)}{q(\pi)q(\theta)q(x_{1:T})}\right] = \mathbb{E}_{q(\pi)}\left[\ln \frac{p(\pi)}{q(\pi)}\right] + \mathbb{E}_{q(\theta)}\left[\ln \frac{p(\theta)}{q(\pi)}\right] \tag{2.4.12}$$

$$+ \mathbb{E}_{q(x_{1:T})q(\pi)q(\theta)}\left[\ln \frac{p(x_{1:T}, y_{1:T}|\pi, \theta)}{q(x_{1:T})}\right]. \tag{2.4.13}$$

For more explicit updates, we assume that each observation prior and likelihood form a conjugate pair of densities and that the prior family is written in natural parameters with the form

$$p(\theta^{(i)}|\eta_\theta^{(i)}) \propto \exp\{\langle \eta_\theta^{(i)}, t_\theta^{(i)}(\theta^{(i)})\rangle\} \qquad p(y|\theta^{(i)}) = \exp\{\langle t_\theta^{(i)}(\theta^{(i)}), (t_y(y), 1)\rangle\}. \tag{2.4.14}$$

Then the update for the factor $q(x_{1:T})$ is

$$q^*(x_{1:T}) \propto \mathbb{E}[\ln p(x_{1:T}, y_{1:T}|\theta, \pi)] \tag{2.4.15}$$

$$= \exp\left\{(\mathbb{E}_{q(\pi)}[\ln \pi^{(0)}])^\mathsf{T} \mathbb{1}_{x_1} + \sum_{t=1}^{T} \mathbb{1}_{x_t}^\mathsf{T} \mathbb{E}_{q(A)}[\ln A] \mathbb{1}_{x_{t+1}} \right.$$

$$\left. + \sum_{t=1}^{T} \langle \mathbb{E}_{q(\theta)}[t_\theta^{(i)}(\theta^{(i)})],\ \mathbb{I}[x_t = i] t_y(y_t, 1)\rangle \right\} \tag{2.4.16}$$

and so, as expected, the optimal factor is also Markov on a chain graph.

For the conjugate updates to $q(\pi)$ and $q(\theta)$, we write the variational factors as

$$q(\theta^{(i)}) \propto \exp\{\langle \widetilde{\eta}_\theta^{(i)}, t_\theta^{(i)}(\theta^{(i)})\rangle\} \qquad q(\pi^{(i)}) = \mathrm{Dir}(\widetilde{\alpha}^{(i)}). \tag{2.4.17}$$

We can compute the expected sufficient statistics over $q(x_{1:T})$ by running the HMM message-passing algorithm. Defining

$$\widetilde{\pi}^{(i)} \triangleq \mathbb{E}_{q(\pi)}[\ln \pi^{(i)}] \qquad \widetilde{L}_{t,i} \triangleq \mathbb{E}_{q(\theta)}[\ln p(y_t|\theta^{(i)})], \tag{2.4.18}$$

and defining $\widetilde{A}$ to be a matrix where the ith row is $\widetilde{\pi}^{(i)}$ for $i = 1, 2, \ldots, N$, we compute

$$\hat{t}_y^{(i)} \triangleq \mathbb{E}_{q(x_{1:T})} \sum_{t=1}^{T} \mathbb{1}[x_t = i] t_y^{(i)}(\bar{y}_t) = \sum_{t=1}^{T} F_{t,i} B_{t,i} \cdot (t_y^{(i)}(\bar{y}_t), 1)/Z \tag{2.4.19}$$

$$(\hat{t}_{\text{trans}}^{(i)})_j \triangleq \mathbb{E}_{q(x_{1:T})} \sum_{t=1}^{T-1} \mathbb{1}[x_t = i, x_{t+1} = j] = \sum_{t=1}^{T-1} F_{t,i} \widetilde{A}_{ij} \widetilde{L}_{t+1,j} B_{t+1,j}/Z \tag{2.4.20}$$

$$(\hat{t}_{\text{init}})_i \triangleq \mathbb{E}_{q(x_{1:T})} \mathbb{1}[x_1 = i] = \widetilde{\pi}_0 B_{1,i}/Z \tag{2.4.21}$$

where $Z = \sum_{i=1}^{N} F_{T,i}$. With these expected statistics, the updates to the parameters of $q(A)$, $q(\pi_0)$, and $q(\theta)$ are then

$$\widetilde{\eta_\theta}^{(i)} \leftarrow \eta_\theta^{(i)} + \hat{t}_y^{(i)} \tag{2.4.22}$$

$$\widetilde{\alpha}^{(i)} \leftarrow \alpha^{(i)} + \hat{t}_{\text{trans}}^{(i)} \tag{2.4.23}$$

$$\widetilde{\alpha}^{(0)} \leftarrow \alpha^{(0)} + \hat{t}_{\text{init}}^{(i)}. \tag{2.4.24}$$

We summarize the overall algorithm in Algorithm 2.6.

■ 2.5 The Dirichlet Process and nonparametric models

The Dirichlet process is used to construct Bayesian nonparametric models, including nonparametric HMMs such that the number of states is unbounded a priori. Bayesian nonparametric methods allow model complexity to be learned flexibly from data and to grow as the amount of data increases. In this section, we review the basic definition of the Dirichlet process and the HDP-HMM.

Definition 2.5.1 (Dirichlet process)**.** *Let $(\Omega, \mathcal{F}, H)$ be a probability space and $\alpha > 0$. We say G is distributed according to a Dirichlet process with parameter αH, and write $G \sim \text{DP}(\alpha H)$ or $G \sim \text{DP}(\alpha, H)$, if $(\Omega, \mathcal{F}, G)$ is a probability space and for every finite partition $\{A_i \subseteq \Omega : i \in [r]\}$ of Ω*

$$\bigcup_{i=1}^{r} A_i = \Omega \qquad i \neq j \implies A_i \cap A_j = \varnothing, \tag{2.5.1}$$

we have

$$(G(A_1), \ldots, G(A_r)) \sim \text{Dir}(\alpha H(A_1), \ldots, \alpha H(A_r)). \tag{2.5.2}$$

By the Kolmogorov consistency theorem, this definition in terms of consistent finite-dimensional marginals defines a unique stochastic process [88]. Though the definition is not constructive, some properties of the Dirichlet process are immediate.

Proposition 2.5.2. *If $G \sim \mathrm{DP}(\alpha H)$, then*

1. *G is* atomic *w.p. 1, meaning it can be written*

$$G = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i} \tag{2.5.3}$$

   for some atoms $\omega_i \in \Omega$ and weights $\pi_i \in (0, 1)$.

2. *If $\theta_i | G \overset{iid}{\sim} G$ for $i = 1, 2, \ldots, N$, then $G|\{\theta_i\}_{i=1}^N$ is distributed as a Dirichlet process with*

$$G|\{\theta_i\}_{i=1}^N \sim \mathrm{DP}(\alpha H + \sum_{i=1}^{N} \delta_{\theta_i}). \tag{2.5.4}$$

Proof. As shown in Ferguson [27], these properties follow from Definition 2.5.1 and finite Dirichlet conjugacy. □

A construction that satisfies Definition 2.5.1 is the *stick breaking process*. In the following, we use $X \sim \mathrm{Beta}(\alpha, \beta)$ to denote that X has the density

$$p(x|\alpha, \beta) \propto x^{\alpha-1}(1-x)^{\beta-1}. \tag{2.5.5}$$

Definition 2.5.3 (Stick-breaking process)**.** *We say $\pi = \{\pi_i : i \in \mathbb{N}\}$ is distributed according to the stick-breaking process with parameter $\alpha > 0$, and write $\pi \sim \mathrm{GEM}(\alpha)$, if*

$$\beta_i \sim \mathrm{Beta}(1, \alpha), \qquad \pi_i = \beta_i \prod_{j<i}(1-\beta_j), \qquad i = 1, 2, \ldots. \tag{2.5.6}$$

Theorem 2.5.4 (Stick-breaking construction)**.** *Let $\pi \sim \mathrm{GEM}(\alpha)$ and $\theta_i \overset{iid}{\sim} H$ for $i \in \mathbb{N}$. If $G = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i}$ then $G \sim \mathrm{DP}(\alpha H)$.*

Proof. See Sethuraman [103]. □

We can define Dirichlet processes that share the same set of atoms and have similar weights using the hierarchical Dirichlet process construction [107]. This construction is useful in defining a Bayesian nonparametric extension of the HMM.

Definition 2.5.5 (Hierarchical Dirichlet process)**.** *We say a collection of random measures $\{G_j : j \in \mathbb{N}\}$ are distributed according to the hierarchical Dirichlet process with parameters α, γ, and H if*

$$G_0 \sim \mathrm{DP}(\alpha, H) \qquad G_j \overset{iid}{\sim} \mathrm{DP}(\gamma, G_0). \tag{2.5.7}$$

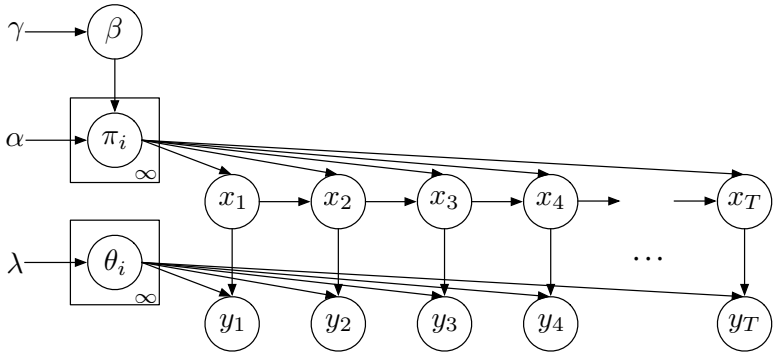

Figure 2.5: Directed graphical model for an HDP-HMM.

We can use the stick-breaking construction of the DP to define a stick-breaking construction of the HDP.

Definition 2.5.6. *We say $(\pi, \theta, x_{1:T}, y_{1:T})$ are distributed according to an HDP-HMM with parameters $\alpha, \gamma > 0$ and base measure H if*

$$\beta \sim \mathrm{GEM}(\gamma), \qquad \pi_i \overset{iid}{\sim} \mathrm{DP}(\alpha\beta), \qquad \theta_i \overset{iid}{\sim} H, \tag{2.5.8}$$

$$x_t \sim \pi_{x_{t-1}}, \qquad y_t \sim p(\,\cdot\,|\theta_{x_t}) \tag{2.5.9}$$

where β is treated as a density with respect to counting measure on $\mathbb{N}$ and where we set $x_1 = 0$. We write $(\pi, \theta, x_{1:T}, y_{1:T}) \sim \mathrm{HDP\text{-}HMM}(\alpha, \gamma, H)$.

Figure 2.5 shows a graphical model for the HDP-HMM.

There are several methods to perform sampling inference in Dirichlet process models. First, exploiting the conjugacy properties of the Dirichlet process, one can analytically marginalize the DP draws, as in the Chinese Restaurant Process (CRP) and Chinese Restaurant Franchise (CRF) samplers for the Dirichlet process and hierarchical Dirichlet process, respectively [107]. However, as before, eliminating variables introduces many dependencies and can result in poor sampler performance for models like the HDP-HMM. One can also work with a finite instantiation of the Dirichlet process draws, so that the sampler only needs to work with the finite Dirichlet marginals, as in the Direct Assignment sampler [107], but such a construction still precludes tree message-passing in an HDP-HMM. The approach we take for most samplers in this thesis is based on approximating a DP prior with a finite symmetric Dirichlet distribution, where the notion of approximation is made precise in the following result.

Theorem 2.5.7 (Weak limit approximation)**.** *Let $(\Omega, \mathcal{F}, H)$ be a probability space, $\alpha > 0$ be a positive constant, and $f : \Omega \to \mathbb{R}$ be any $(\mathcal{F}, \mathcal{B}(\mathbb{R}))$-measurable function.*

Consider the finite model of size K given by

$$\theta_i \overset{iid}{\sim} H \qquad \pi = (\pi_1, \ldots, \pi_K) \sim \mathrm{Dir}(\alpha_1/K, \cdots, \alpha_K/K) \tag{2.5.10}$$

and define the measure $G^K = \sum_{i=1}^K \pi_i \delta_{\theta_i}$. Then as $K \to \infty$ we have

$$\int f(\omega) G^K(d\omega) \xrightarrow{\mathcal{D}} \int f(\omega) G(d\omega) \tag{2.5.11}$$

where $G \sim \mathrm{DP}(\alpha H)$.

Proof. See Ishwaran and Zarepour [57, Theorem 2], which also gives rates of convergence and bounds on the probabilities of some error events. □

Based on this approximation result, we can define Bayesian nonparametric models and perform approximate inference with finite models of size K, where K becomes an algorithm parameter rather than a model parameter. With these finite approximations we can exploit graphical model structure and tree message-passing algorithms in both Gibbs sampling and mean field algorithms for time series models defined with the HDP.

Algorithm 2.2 HMM Forwards Messages

Input: transition potentials A, emission potentials L, initial state potential $\pi^{(0)}$
Output: HMM forward messages F
  function HMMFwdMessages(A,L,$\pi^{(0)}$)
    $F_{1,:} \leftarrow \pi^{(0)} \odot L_{1,:}$
    for $t = 2, 3, \ldots, T$ **do**
      $F_{ti} \leftarrow \sum_{j=1}^{N} F_{t-1,j} A_{ji} L_{ti}$
    return F

Algorithm 2.3 HMM Backward Messages

Input: transition potentials A, emission potentials L
Output: HMM backwards messages B
  function HMMBwdMessages(A,L)
    $B_{T,:} \leftarrow 1$
    for $t = T-1, T-2, \ldots, 1$ **do**
      $B_{t,i} \leftarrow \sum_{j=1}^{N} A_{ij} B_{t+1,j} L_{t+1,j}$
    return B

Algorithm 2.4 Bayesian HMM Gibbs sampling

Input: α, λ, $\bar{y}_{1:T}$
Output: samples $\{(\hat{x}_{1:T}, \hat{\theta}, \hat{\pi})^{(t)}\}$
  Initialize $x_{1:T}$
  for $t = 1, 2, \ldots$ **do**
    for $i = 1, 2, \ldots, N$ **do**
      $\pi^{(i)} \leftarrow$ sample $\mathrm{Dir}(\alpha^{(i)} + n_{i:})$ with $n_{ij} = \sum_t \mathbb{I}[x_t = i, x_{t+1} = j]$
      $\theta^{(i)} \leftarrow$ sample $p(\theta | \lambda, \{y_t : x_t = i\})$
    $\pi^{(0)} \leftarrow$ sample $\mathrm{Dir}(\alpha^{(0)} + \mathbb{1}_{x_1})$
    $x_{1:T} \leftarrow$ HMMSampleStates($\pi^{(0)}, A, L$) with $L_{t,i} = p(y_t | \theta^{(i)})$
    $(\hat{x}_{1:T}, \hat{\theta}, \hat{\pi})^{(t)} \leftarrow (x_{1:T}, \theta, \pi)$

Algorithm 2.5 HMM state sequence sampling

Input: $\pi^{(0)}$, A, L
Output: a sample $x_{1:T}$
  function HMMSampleStates(A,L)
    $B \leftarrow$ HMMBwdMessages(A, L)
    $x_1 \leftarrow$ sample $\pi_i^{(0)} B_{1,i} L_{1,i}$ over i
    for $t = 2, 3, \ldots, T$ **do**
      $x_t \leftarrow$ sample $A_{x_{t-1},i} B_{t,i} L_{t,i}$ over i

Algorithm 2.6 HMM Mean Field

Initialize variational parameters $\widetilde{\eta}_\theta^{(i)}$, $\widetilde{\alpha}^{(i)}$, $\widetilde{\alpha}^{(0)}$
for $t = 1, 2, \ldots$ until convergence **do**
  $F \leftarrow \text{HMMFwdMessages}(\widetilde{A}, \widetilde{L}, \widetilde{\pi}^{(0)})$
  $B \leftarrow \text{HMMBwdMessages}(\widetilde{A}, \widetilde{L})$
  Using F and B, compute each $\hat{t}_y^{(i)}$, $\hat{t}_{\text{trans}}^{(i)}$, and $\hat{t}_{\text{init}}$
    with Eqs. (2.4.19)-(2.4.21)
  Update $\widetilde{\eta}_\theta^{(i)}$, $\widetilde{\alpha}^{(i)}$, $\widetilde{\alpha}^{(0)}$ for $i = 1, 2, \ldots, N$
    with Eqs. (2.4.22)-(2.4.24)