Chapter 2

Background

This chapter outlines the background necessary for subsequent developments in this thesis. Graphical models provide a flexible framework for specifying globally consistent probability models based on local constraints. The primary focus of this thesis is problems that arise in using such models. We begin in Section 2.1 with an introduction to the basics of graphical models, and the relevant problems of inference and estimation. As a prelude to introducing graphical models, this section also contains a brief primer on graph theory. Section 2.2 introduces a particular representation of distributions defined by graphical models — namely, the exponential family. Associated with such families are a variety of elegant results, known collectively as information geometry. A third concept central to this thesis is that of a variational formulation. Accordingly, we devote Section 2.3 to an overview of variational methods, with a particular emphasis on mean field theory.

■ 2.1 Graphical models

Graphical models are a powerful framework for representing and manipulating probability distributions over sets of random variables. Indeed, stochastic processes defined on graphs arise in a variety of fields, including coding theory [72], statistical physics [15,32,136], artificial intelligence [138], computer vision [67], system theory [14] and statistical image processing [127]. The power of graphical models derives from the correspondence that they establish between the probabilistic concept of *conditional independence*, and the graph-theoretic notion of *node separation*.

We begin in Section 2.1.1 with a brief but self-contained introduction to the basics of graph theory. There are many books available to the reader interested in more background on graph theory [e.g., 16,22,26,27]. In Section 2.1.2, we turn to the basics of graphical models. More background on graphical models can be found in the books [69, 102,105,122]; another helpful source is the edited collection of papers [104]. Section 2.1.3 introduces the problem of estimation or inference in graphical models, which is central to many parts of this thesis. Section 2.1.4 briefly discusses exact inference algorithms for tree-structured graphs; more details can be found in Appendix A. In Section 2.1.5, we describe the notion of a junction tree, which is important both in a purely graph-theoretic context and for the purposes of inference in graphical models.

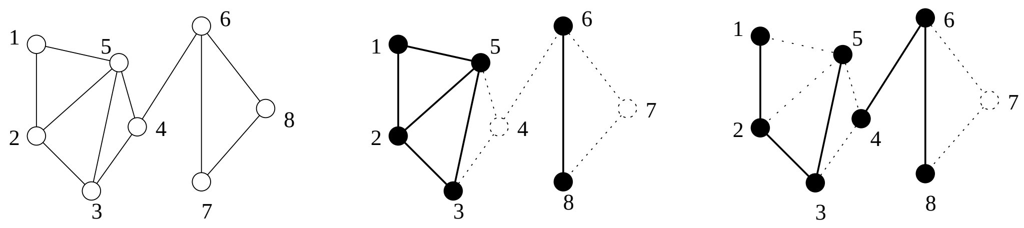
2.1.1 Basics of graph theory

A graph $G = (V, E)$ consists of a set of nodes or vertices $V = \{1, \dots, N\}$ that are joined by a set of edges E. Edges in a graph can either be directed or undirected; this thesis will focus exclusively on undirected graphs. For an undirected graph, the notation (s, t) (or equivalently, (t, s)) denotes an undirected edge between nodes s and t in the vertex set. Throughout this thesis, we will focus on simple graphs, for which self-loops (i.e., an edge from a node back to itself), and multiple edges between the same pair of nodes are both forbidden.

A subgraph $H = (V(H), E(H))$ of a graph G is formed by a particular subset $V(H)$ of the vertices, and a particular subset $E(H)$ of the edges of G. It is often convenient to consider subgraphs induced by particular subsets of the vertex set, or by particular subsets of the edge set. First of all, given a subset S of the vertex set V, the subgraph induced by S is given by $G[S] = (S, E[S])$ where $E[S] = \{ (s,t) \in E \mid s,t \in S \}$. The graph $G[S]$ is called a *node-induced subgraph*. Similarly, given a subset $F \subset E$ of the edge set, the subgraph induced by F is given by $G(F) = (V(F), F)$, where

$$V(F) \triangleq \{ u \in V \mid \exists\, v \in V \text{ s.t } (u,v) \in F \}$$

This graph $G(F)$ is called an *edge-induced* subgraph. Examples of node and edge-induced subgraphs are given in Figure 2.1.



(a) (b) (c)

Figure 2.1. Illustration of node and edge-induced subgraphs. Vertices and edges in the subgraph are shown in dark circles and solid lines (respectively), while those not in the subgraph are shown with dotted open circles and dotted lines (respectively). (a) Graph G with cycles. (b) The node-induced subgraph $G[S]$ for $S = \{1, 2, 3, 5, 6, 8\}$. (c) The edge-induced subgraph $G(F)$ with $F = \{(1, 2), (2, 3), (3, 5), (4, 6), (6, 8)\}$.

For any $s \in V$, the set of *neighbors* of s in G is given by

$$\mathcal{N}(s) \triangleq \{ t \in V \mid (s,t) \in E \} \tag{2.1}$$

The *degree* of a node s, denoted $d(s)$, corresponds to the number of neighbors (i.e., the cardinality $|\mathcal{N}(s)|$ of the neighbor set).

A *path* is a graph P consisting of the vertex set $V(P) = \{s_0, s_1, \dots, s_k\}$ and a set of distinct edges $E(P) = \{(s_0, s_1), \dots (s_{k-1}, s_k)\ \}$. The vertices s_0 and s_k are the end vertices of the path, and $l(P) = k$ is the length of the path. We say that P is a path from s_0

to s_k. A *cycle* is a path from a node s back to itself formed of a sequence of distinct edges. I.e., a cycle consists of a sequence of distinct edges $\{ (s_1, s_2), (s_2, s_3), \ldots, (s_{k-1}, s_k) \}$ such that $s_1 = s_k$. We say that a graph is *connected* if for each pair $\{s, t\}$ of distinct vertices, there is a path from s to t. A *component* of the graph is a maximal connected subgraph. The notation $c(G)$ denotes the number of (connected) components in the graph G.

Of particular interest are graphs without cycles:

Definition 2.1.1 (Forests and trees). A *tree* $\mathcal{T}$ is a cycle-free graph consisting of a single connected component. A *forest* is formed by the union of a collection of trees. Given a graph G, a *spanning tree* is an embedded tree (i.e., a tree-structured subgraph of G) that reaches each vertex. See Figure 2.2. for illustration of these concepts.

Definition 2.1.2 (Cliques). A *clique* of a graph G is any fully connected subset of the vertex set V. A clique is *maximal* if it is not properly contained within any other clique.

Figure 2.3 illustrates the structure of cliques of sizes one through four. Note that any single node is itself a clique, but not a maximal clique unless it has no neighbors. If we return to Figure 2.1(a), nodes $\{1, 2, 5\}$ form a 3-clique, but nodes $\{1, 2, 5, 3\}$ do *not* form a 4-clique, since node 1 is not connected (directly) to node 3.

Let $\mathbf{C} = \mathbf{C}(G)$ denote the set of all cliques in a graph G. For instance, given a tree $\mathcal{T}$, the clique set $\mathbf{C}(\mathcal{T})$ consists of the union $V \cup E$ of the vertex set with the edge set. We use $\mathcal{C}$ to denote an arbitrary member of $\mathbf{C}$ (i.e., a particular clique of G). Given a subset of the clique set $\mathbf{C}$, it is natural to define the following generalization of an edge-induced subgraph:

Definition 2.1.3 (Clique-induced subgraphs). Given a subset $\mathbf{B} \subset \mathbf{C}$ of the clique set, let $G(\mathbf{B})$ denote the subgraph of G induced by the cliques in $\mathbf{B}$. More precisely, $G(\mathbf{B}) = (V(\mathbf{B}); E(\mathbf{B}))$ where

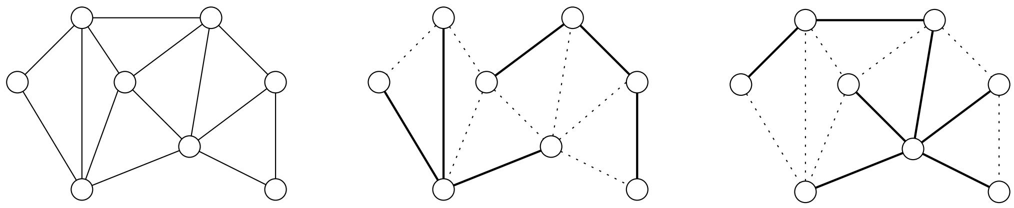
$$V(\mathbf{B}) \triangleq \{ s \in V \mid s \in \mathcal{C} \text{ for some } \mathcal{C} \in \mathbf{B} \} \tag{2.2a}$$

$$E(\mathbf{B}) \triangleq \{ (s, t) \in E \mid s, t \in \mathcal{C} \text{ for some } \mathcal{C} \in \mathbf{B} \} \tag{2.2b}$$

Note the clique set of $G(\mathbf{B})$ can be strictly larger than $\mathbf{B}$. For example, if we consider a single cycle G on three nodes with $\mathbf{B} = \{(1, 2), (2, 3), (1, 3)\}$, then the clique set of $G(\mathbf{B})$ includes the 3-clique $\{1, 2, 3\} \notin \mathbf{B}$.

The notion of triangulation will play a central role in the junction tree representation of graphical models, to be discussed in Section 2.1.5.

Definition 2.1.4 (Triangulated). A graph G is *triangulated* if every cycle of length 4 or greater has a chord (i.e., an edge joining two vertices not adjacent in the cycle). See Figure 2.4 for illustrations of triangulated versus non-triangulated graphs.



(a) (b) (c)

Figure 2.2. (a) Graph G with cycles. (b) A forest embedded within G. (c) Embedded spanning tree that reaches each vertex of G.

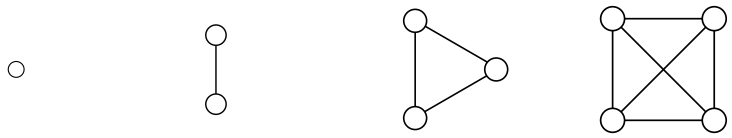
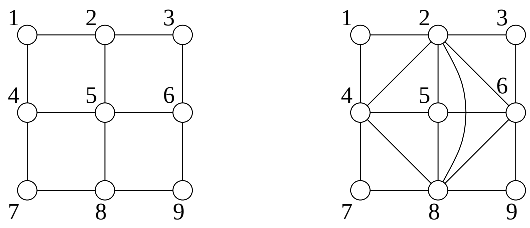


Figure 2.3: Graph cliques of size 1 through 4.



(a) (b)

Figure 2.4. Illustration of a non-triangulated versus triangulated graph. (a) This 3×3 grid is not triangulated; it has many four cycles (e.g., the cycle formed by nodes $1-2-5-4-1$) that lack a chord. (b) Here is one triangulated version of the 3×3 grid, formed by adding the extra edges $\{(2,4),(4,8),(2,6),(6,8),(2,8)\}$. The extra edge $(2,8)$ is added as a chord for the 4-cycle formed by nodes $2-4-8-6-2$.

Given a graph that is not triangulated, it is always possible to form a triangulated version $\widetilde{G}$ by adding chords to cycles as necessary. However, this triangulated version need not be unique; that is, a given untriangulated graph G may have a number of possible triangulations.

It is useful to distinguish vertices (and edges), that if removed from the graph, increase the number of connected components:

Definition 2.1.5 (Cut vertices and bridges). A vertex is a *cut vertex* if its deletion from the graph increases the number of connected components. A *bridge* is an edge whose deletion increases the number of connected components. (See Figure 2.5).

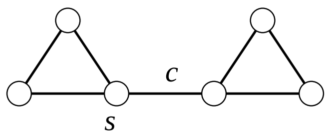


Figure 2.5: Vertex s is a cut vertex in the graph shown, whereas edge c is a bridge.

2.1.2 Basics of graphical models

Given a graph $G = (V, E)$, a probabilistic graphical model is formed by associating with each node $s \in V$ a random variable x_s taking values in the sample space $\mathcal{X}$. This sample space can be either a continuum (e.g., $\mathcal{X} = \mathbb{R}$), or the discrete alphabet $\mathcal{X} = \{0, \ldots, m-1\}$. In this latter discrete case, the underlying sample space $\mathcal{X}^N$ is the set of all N vectors $\mathbf{x} = \{x_s \mid s \in V\}$ over m symbols, so that $|\mathcal{X}^N| = m^N$.

In a graphical model, the edges of the underlying graph represent probabilistic dependencies between variables, and come in two varieties — directed or undirected. Although the probabilistic interpretation of directed and undirected edges is different, any directed graph can be converted to an equivalent[1] undirected graph [see, e.g., 138]. In this thesis, we restrict our attention to undirected graphs.

The stochastic processes of interest are those which are Markov with respect to the underlying graph. To define this concept, let $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ be subsets of the vertex set V. Let $\mathbf{x}_{\mathcal{A}|\mathcal{B}}$ be the random variables in $\mathcal{A}$ conditioned on those in $\mathcal{B}$. The set $\mathcal{B}$ separates $\mathcal{A}$ and $\mathcal{C}$ if in the modified graph with $\mathcal{B}$ removed, there are no paths between nodes in the sets $\mathcal{A}$ and $\mathcal{C}$ (see Figure 2.6).

Definition 2.1.6. A stochastic process $\mathbf{x}$ is *Markov* with respect to the graph G if $\mathbf{x}_{\mathcal{A}|\mathcal{B}}$ and $\mathbf{x}_{\mathcal{C}|\mathcal{B}}$ are conditionally independent whenever $\mathcal{B}$ separates $\mathcal{A}$ and $\mathcal{C}$.

[1] However, it may no longer be possible to read directly certain conditional independencies from the undirected graph.

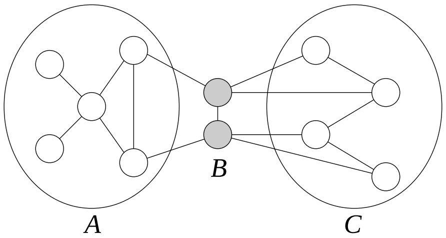


Figure 2.6. Illustration of the relation between conditional independence and graph separation. Here the set of nodes $\mathcal{B}$ separates $\mathcal{A}$ and $\mathcal{C}$, so that $\mathbf{x}_{\mathcal{A}|\mathcal{B}}$ and $\mathbf{x}_{\mathcal{C}|\mathcal{B}}$ are conditionally independent.

This definition of Markovianity constitutes a generalization of the concept as applied to a discrete time series. Indeed, a time series sampled at discrete instants can be viewed as a stochastic process defined on a chain. For such a graph, Definition 2.1.6 corresponds to the usual notion that the past and future are conditionally independent given the present.

A graph strongly constrains the distribution of a Markov process. Indeed, the *Hammersley-Clifford theorem* [21,80] guarantees that distributions of Markov processes over graphs can be expressed in factorized form as products of so-called *compatibility functions* defined over the cliques:

Theorem 2.1.1 (Hammersley-Clifford). *Let G be a graph with a set of cliques $\mathbf{C}$. Suppose that a distribution*[2] *p over a discrete random vector $\mathbf{x}$ is formed as a normalized product of nonnegative functions over the cliques:*

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{\mathcal{C} \in \mathbf{C}} \psi_{\mathcal{C}}(\mathbf{x}) \tag{2.3}$$

where $\psi_{\mathcal{C}}(\mathbf{x})$ is a compatibility function depending only on the subvector $\mathbf{x}_{\mathcal{C}} = \{x_s \mid s \in \mathcal{C}\}$; and $Z \triangleq \sum_{\mathbf{x}} \prod_{\mathcal{C} \in \mathbf{C}} \psi_{\mathcal{C}}(\mathbf{x})$ is the partition function. Then the underlying process $\mathbf{x}$ is Markov with respect to the graph. Conversely, the distribution p of any Markov random field over G that is strictly positive (i.e., $p(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}^N$) can be represented in this factorized form.

Remarks: There a variety of proofs of this result [e.g., 21, 80]; see Clifford [39] for a historical overview. One of the most elegant proofs [80] uses the Möbius inversion formula [see, e.g., 29]. Note that this theorem generalizes the usual factorizations of Markov chains, for which the compatibility functions are formed by forward (or backward) transition functions defined on the edges (i.e., maximal cliques for a chain). See Lauritzen [122] for an example of a non-positive distribution (i.e., $p(\mathbf{x}) = 0$ for some $\mathbf{x} \in \mathcal{X}^N$) for which the converse is false.

[2]Strictly speaking, p is a probability mass function for discrete random variables; however, we will use distribution to mean the same thing.

2.1.3 State estimation or inference

A problem that arises in many applications of interest is that of estimating the random vector $\mathbf{x} = \{x_s \mid s \in V\}$ based on a set of noisy observations $\mathbf{y} = \{y_s \mid s \in V\}$. For instance, in image processing or computer vision [67,127], the vector $\mathbf{x}$ could represent an image defined on a grid, and $\mathbf{y}$ could represent a noisy or blurred version of this image. Similarly, in the context of channel coding [69,72], the vector $\mathbf{x}$ would represent message bits, whereas $\mathbf{y}$ would correspond to the received bits.

In all cases, the goal is to *estimate* or to draw statistical *inferences* about the unobserved $\mathbf{x}$ based on the observations $\mathbf{y}$. The observation model can be formulated mathematically in the form of a conditional distribution. In particular, we assume that for each node $s \in V$, the variable y_s is a noisy observation of x_s, specified by the conditional density $p(y_s|x_s)$. We assume that the observations $\mathbf{y}$ are conditionally independent given the hidden variables[3] $\mathbf{x}$, so that $p(\mathbf{y}|\mathbf{x}) = \prod_{s \in V} p(y_s|x_s)$.

Of central interest for problems of estimation or inference is the posterior density $p(\mathbf{x}|\mathbf{y})$, which defines a variety of estimators:

1. The *maximum a posteriori* (MAP) estimate corresponds to the peak or mode of the posterior density — that is: $\widehat{\mathbf{x}}_{MAP} = \arg\max_{\mathbf{x} \in \mathcal{X}^N} p(\mathbf{x}|\mathbf{y})$

2. Also of interest are posterior marginals of a subset of variables. For instance, for a discrete process $\mathbf{x}$, the single node marginals are given by

$$p(x_s|\mathbf{y}) = \sum_{\{\mathbf{x}' \mid x'_s = x_s\}} p(\mathbf{x}'|\mathbf{y}) \tag{2.4}$$

   Here the notation means summing over all configurations $\mathbf{x}' \in \mathcal{X}^N$ such that $x'_s = x_s$. For a continuous-valued process, this summation should be replaced by integration.

By combining the prior in equation (2.3) with the observation density via Bayes rule, we have:

$$p(\mathbf{x} \mid \mathbf{y}) = \frac{1}{Z} \prod_{C \in \mathbf{C}} \psi_C(\mathbf{x}) \prod_s p(y_s|x_s) \tag{2.5}$$

Note that each individual node forms a singleton clique, meaning that some of the factors in (2.3) may involve functions of each individual variable. As a consequence, the transformation from the prior distribution $p(\mathbf{x})$ of equation (2.3) to the conditional distribution $p(\mathbf{x} \mid \mathbf{y})$ of equation (2.5) is simply to modify the singleton factors of equation (2.3). As a result, from here onwards, we suppress explicit mention of measurements, since problems of estimation or inference for either $p(\mathbf{x})$ or $p(\mathbf{x} \mid \mathbf{y})$ are of identical structure and complexity.

[3]This assumption entails no loss of generality, since any observation that is a function of variables at multiple nodes can be merged into a clique potential that includes those nodes.

The computations of the MAP estimate or of the single-node marginals are both well-defined tasks. The latter task will be of primary interest in this thesis. Difficulties arise from different sources, depending on whether $\mathbf{x}$ is a discrete or continuous-valued process. For a continuous process, it may not be possible to evaluate analytically the necessary integrals. This difficulty is relevant even for small problems. For a discrete process, on the other hand, computing a marginal simply involves a discrete summation, which is a straightforward operation for small problems. Here the difficulty arises as the problem size grows. In particular, given a discrete-valued process on N nodes with $m \geq 2$ states, the number of terms in the summation of equation (2.4) explodes exponentially as m^{N-1}. Consequently, for sufficiently large graphs, it will be impossible to perform the discrete summation. A similar curse of dimensionality applies to the computation of the MAP estimate.

■ 2.1.4 Exact inference in trees

For a Markov chain, there exist highly efficient algorithms for computing the MAP estimate, or the single-node marginals at each node. These algorithms exploit the Markov properties of a chain — namely, that the past and future are conditionally independent given the present — to perform the necessary computations in a recursive and hence efficient manner. For the linear-Gaussian problem, this formulation leads to the Rauch-Tung-Striebel smoother [146]. For a discrete-state hidden Markov chain, the resulting algorithm is known as the $\alpha - \beta$ algorithm in the speech processing literature [143].

Interestingly, these recursive algorithms can be generalized to trees, which are singly-connected graphs without cycles. (A chain is a special case of a tree.) An important property of trees is that their nodes can be assigned a partial ordering in terms of their depth in relation to an arbitrary node designated as the root. That is, the root is scale 0; the immediate descendants (i.e., children) of the root are scale 1; and so on down to the leaves (terminal nodes) of the tree. With this partial ordering, the most efficient implementation of a tree inference algorithm follows a two-pass form, first sweeping up from the leaves to the root, and then downwards from the root to the leaves. For a discrete process, the computational complexity of these algorithms is $\mathcal{O}(m^2 N)$. See Appendix A for more details about such tree algorithms.

■ 2.1.5 Junction tree representation

The set of cliques of a Markov chain are single nodes and pairs of adjacent nodes. In this case, the compatibility functions $\{\psi_{\mathcal{C}}\}$ of equation (2.3) can always be written as a function of local marginal and conditional distributions. For example, the standard forward factorization of a Markov chain on three nodes is in terms of an initial distribution and transitions:

$$p(\mathbf{x}) = p(x_1)\, p(x_2 \,|\, x_1)\, p(x_3 \,|\, x_2)$$

There is an alternative factorization that is symmetric with respect to the nodes — namely $p(\mathbf{x}) = [p(x_1, x_2)/p(x_1)p(x_2)]\, [p(x_2, x_3)/p(x_2)p(x_3)]p(x_1)p(x_2)p(x_3)$. More generally, the same kind of symmetric factorization holds for any tree-structured graph $\mathcal{T}$:

$$p(\mathbf{x}) = \prod_{s \in V} p(x_s) \prod_{(s,t) \in E} \frac{p(x_s, x_t)}{p(x_s)p(x_t)} \tag{2.6}$$

That is, for a tree, the compatibility functions of equation (2.3) can always be represented directly in terms of local marginal distributions: $\psi_s(x_s) = p(x_s)$ for each node $s \in V$; and $\psi_{st}(x_s, x_t) = [p(x_s, x_t)/p(x_s)p(x_t)]$ for each edge $(s,t) \in E$.

In contrast, for a graph with cycles, the compatibility functions do not, in general, have any direct correspondence with local marginal distributions on those same cliques.[4] However, such a correspondence does hold on a graph formed of suitably aggregated nodes, which is the subject of the *junction tree representation*. The basic idea is to cluster nodes within the original graph G so as to form a *clique tree* — that is, an acyclic graph whose nodes are formed by cliques of G. We use the calligraphic $\mathcal{C}$ to refer to a given node of the clique tree (i.e., a given clique of G).

Having formed a tree, it is tempting to simply apply a standard tree inference algorithm. However, the clique tree must satisfy an additional restriction so as to ensure consistency of probabilistic inference on the tree. To understand the source of this problem, consider the single loop on 4 nodes shown in Figure 2.7(a), as well as the clique tree (one of many possible) shown in Figure 2.7(b). Here ellipses represent nodes of the clique tree (i.e., cliques of the original graph), whereas the boxes represent *separator sets*, which correspond to intersections of nodes adjacent on the clique tree. Observe that node 3 occurs twice in the clique tree, once in each of the cliques $\{1, 3\}$

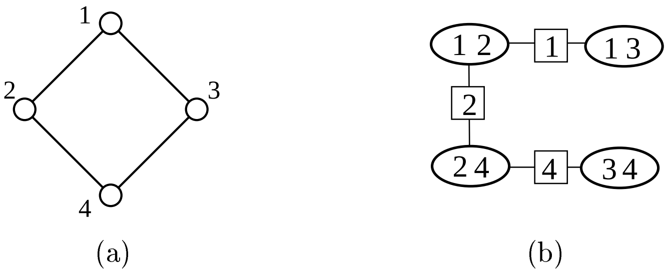


Figure 2.7. A simple example showing the necessity of the running intersection property for probabilistic consistency. (a) Single loop on 4 nodes. (b) One possible clique tree for the graph in (a). This clique tree fails the running intersection property.

and $\{3, 4\}$. However, any tree inference algorithm applied to the clique tree of (b) will not enforce the implicit constraint that the corresponding random variable x_3 in clique $\{1, 3\}$ must match the x_3 in clique $\{3, 4\}$. As a result, running a tree inference algorithm on the graph in (b) will not yield the correct results for the single loop of (a).

[4]The simplest example to consider is the single cycle on 4 nodes; here the pairwise compatibility functions can never correspond to $P_{st}/P_s P_t$. See Proposition 5.4.4 in Chapter 5.

What is required is a mechanism for enforcing consistency among the different appearances of the same random variable. Note that the same problem does not arise for node 2, although it also appears in both of the two cliques $\{1,2\}$ and $\{2,4\}$. The difference is that node 2 also appears in all separator sets in the path between these two cliques, which provides a pipeline for transmitting and enforcing the associated consistency constraints. This motivates the following definition:

Definition 2.1.7. A clique tree has the *running intersection property* if for any two clique nodes $\mathcal{C}_1$ and $\mathcal{C}_2$, all nodes on the unique path joining them contain the intersection $\mathcal{C}_1 \cap \mathcal{C}_2$. A clique tree with this property is known as a *junction tree.*

For what type of graphs can one build junction trees? It is clear that no clique tree of the single loop in Figure 2.7(a) has the running intersection property. (Since the clique tree of Figure 2.7(b) does not satisfy running intersection, by a symmetry argument neither can any other clique tree.) An important result in graph theory establishes a correspondence between junction trees and triangulated graphs (see Definition 2.1.4).

Proposition 2.1.1. *A graph G has a junction tree $\Longleftrightarrow$ it is triangulated.*

Proof. See Lauritzen [122]. □

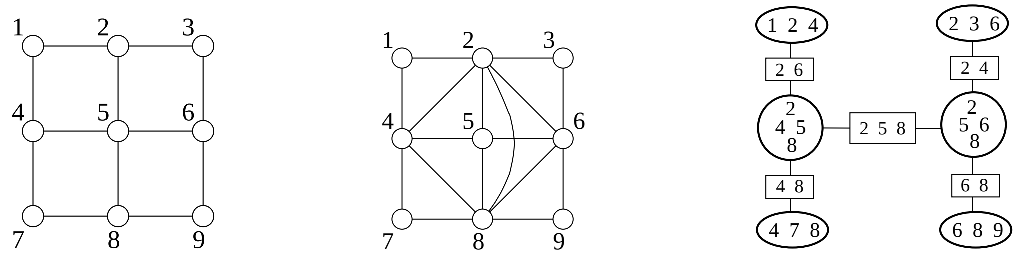
This proposition leads to a method for exact inference on arbitrary graphs:

Algorithm 2.1.1 (Junction tree).

1. Given a graph with cycles G, triangulate it by adding edges as necessary.
2. Form a junction tree associated with the triangulated graph $\widetilde{G}$.
3. Run a tree inference algorithm on the junction tree.

Although this procedure is sound in principle, its practical use is limited. For most applications of interest, the size of the cliques in the triangulated version $\widetilde{G}$ grows with problem size. As a result, the state cardinality of the supernodes in the junction tree grows exponentially, meaning that applying tree algorithms rapidly becomes prohibitively complex. This explosion in the state cardinality is another demonstration of the intrinsic complexity of exact computations for graphs with cycles.

Example 2.1.1. To illustrate the junction tree procedure and its associated complexities, we consider the 3×3 grid shown in Figure 2.8(a). The first step is to form a triangulated version $\widetilde{G}$, as shown in Figure 2.8(b). Note that the graph would not be triangulated if the additional edge joining nodes 2 and 8 (shown in a dashed line) were not present. Without this edge, the 4-cycle $(2-4-8-6-2)$ would lack a chord. As a result of this additional edge, the junction tree has two 4-cliques in the middle, as



(a) (b) (c)

Figure 2.8. Illustration of junction tree procedure. (a) Original graph is a 3×3 grid. (b) Triangulated version of original graph. Note the two 4-cliques in the middle. (c) Corresponding junction tree for triangulated graph in (b), with maximal cliques depicted within ellipses, and separator sets within rectangles.

shown in Figure 2.8(c). Consequently, running a tree inference algorithm on the junction tree involves dealing with variables with state cardinalities of m^4. This difficulty only worsens as the grid size grows.

Despite its limited practical use, the junction tree procedure provides conceptual insight into the inherent complexity of a given distribution on a graph. In particular, it gives rise to an alternative representation of the distribution, in terms of local marginal distributions on maximal cliques and separator sets. That is,

$$p(\mathbf{x}) = \frac{\prod_{\mathcal{C} \in \mathbf{C}} p(\mathbf{x}_{\mathcal{C}})}{\prod_{\mathcal{S} \in \mathbf{S}} p(\mathbf{x}_{\mathcal{S}})} \tag{2.7}$$

where $\mathbf{C}$ is the set of all maximal cliques of $\widetilde{G}$, and $\mathbf{S}$ is the associated set of separators. Unlike the representation of equation (2.3), equation (2.7) provides a decomposition directly in terms of local marginal distributions. The price to be paid is that the decomposition involves functions defined over larger clusters of variables. Note that equation (2.6) is a particular case of this decomposition, where the maximal cliques are the edges of the ordinary tree, and the separator sets correspond to nodes with degree greater than one.

■ 2.2 Exponential families and information geometry

Exponential families of distributions and their associated geometry have been studied extensively in applied probability theory and statistics. Work in this area dates back to Rao [145] in 1945, who developed the geometric role of the Fisher information matrix. Subsequent contributions were made by a variety of people, including Chentsov [33,34], Csiszár [44–47], Barndorff-Nielson [13] and Amari [5–7]. This section contains a brief introduction to this body of theory, which is often referred to as *information geometry*. We emphasize only those concepts necessary for our subsequent development; see the references above, or the edited collection of papers in [83] for further details. Although information geometry applies to any exponential family of distributions, we focus here on such distributions in the specific context of graphical models.

2.2.1 Exponential representations

Equation (2.3) decomposes a graph distribution as a product of compatibility functions defined on the cliques. A related representation is the Gibbs form, in which a distribution is specified as the exponential of a sum of functions on the cliques. In the context of graphical models, an exponential family constitutes a collection of such Gibbs distributions:

$$p(\mathbf{x};\theta) = \exp\big\{\sum_{\alpha}\theta_\alpha\phi_\alpha(\mathbf{x}) - \Phi(\theta)\big\} \tag{2.8a}$$

$$\Phi(\theta) = \log\big(\sum_{\mathbf{x}\in\mathcal{X}^N}\exp\{\sum_{\alpha}\theta_\alpha\phi_\alpha(\mathbf{x})\}\big) \tag{2.8b}$$

The quantity Φ defined in equation (2.8b) is the *log partition function* that serves to normalize the distribution; when the sample space $\mathcal{X}^N$ is continuous, the summation defining Φ should be replaced by an integral.

Any exponential family is specified by a collection of potential functions $\{\phi_\alpha \mid \alpha \in \mathcal{A}\}$, where $\mathcal{A}$ is a finite index set. The domain of the *exponential parameter vector* θ is the set

$$\Theta \triangleq \{\,\theta \in \mathbb{R}^{|\mathcal{A}|} \mid \Phi(\theta) < \infty\}$$

In the discrete case, this imposes no restrictions (i.e., $\Theta = \mathbb{R}^{|\mathcal{A}|}$); in continuous examples, Θ can be a strict subset of $\mathbb{R}^{|\mathcal{A}|}$. In this thesis, we focus primarily on the discrete case.

Each parameter vector $\theta \in \Theta$ indexes a particular member $p(\mathbf{x};\theta)$ of the family, assuming that the set of clique potentials $\boldsymbol{\phi} = \{\phi_\alpha\}$ is fixed. With some abuse of notation, we will often use the parameter vector θ itself as a shorthand for the associated distribution.

Minimal representations

It is typical to define an exponential family with a collection of functions $\boldsymbol{\phi} = \{\phi_\alpha\}$ that are linearly independent. Indeed, if there were any linear dependencies, they could be eliminated without sacrificing any expressive power of the exponential model. This condition gives rise to a so-called *minimal representation* [e.g., 13], in which there is a unique parameter vector θ associated with each distribution. In this case, the *dimension* of the exponential family, denoted by $d(\theta)$, is given by $|\mathcal{A}|$.

To illustrate these definitions, we consider some simple examples:

Example 2.2.1. Consider a scalar Gaussian random variable $x \sim \mathcal{N}(\mu, \sigma^2)$. Then its density has the exponential representation

$$p(x; \theta) = \exp\{\theta_1 x + \theta_2 x^2 - \Phi(\theta)\} \tag{2.9}$$

I.e. here we have $\phi_1(x) = x$ and $\phi_2(x) = x^2$. By completing the square, we obtain relations between the exponential parameters (θ_1, θ_2) and the mean and variance — namely, $\theta_2 = -1/[2\sigma^2]$ and $\theta_1 = \mu/\sigma^2$. Here the dimension of the family is $d(\theta) = 2$. Moreover, the domain of θ is the half plane

$$\Theta = \{(\theta_1, \theta_2) \in \mathbb{R}^2 \mid \theta_2 < 0\}$$

The restriction on θ_2 is required so that the associated integral defining the log partition function — namely, $\Phi(\theta) = \int_{-\infty}^{\infty} \exp\{\theta_1 x + \theta_2 x^2\} dx$ — is finite.

Example 2.2.2. Now consider a binary process (i.e., $\mathbf{x} \in \{0, 1\}^N$) defined on a graph with pairwise maximal cliques (i.e., whose maximal cliques are pairs of nodes). The standard (minimal) representation corresponds to the *Boltzmann machine* [e.g., 106], also known as the *Ising model* in statistical physics [15, 32]:

$$p(\mathbf{x}; \theta) = \exp\Big\{\sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t - \Phi(\theta)\Big\} \tag{2.10}$$

where θ_{st} is the strength of edge (s, t), and θ_s is the node parameter for node s. In this case, $d(\theta) = |V| + |E| = N + |E|$, and the domain Θ of θ is all of $\mathbb{R}^{d(\theta)}$.

Examples 2.2.1 and 2.2.2 illustrate that the sample space $\mathcal{X}$ is critical in assessing the linear independence of a set of functions $\{\phi_\alpha\}$, and hence the minimality of the representation. In Example 2.2.1, the functions x and x^2 are linearly independent over $\mathbb{R}$, so that equation (2.9) constitutes a minimal representation. In contrast, these same functions are *not* linearly independent over $\{0, 1\}$.

Example 2.2.3. We now consider an extension of Example 2.2.2, with $\mathbf{x} \in \{0, 1\}^N$. The Ising model corresponds to pairwise maximal cliques. To incorporate higher-order cliques (e.g., the 3-clique $\{s, t, u\}$), we add a multinomial of the form $x_s x_t x_u$, with corresponding exponential parameter θ_{stu}. Cliques of higher order are incorporated in a similar fashion, so that the minimal representation of the most general distribution (i.e., possibly on the complete graph) is of the form:

$$\begin{aligned} p(\mathbf{x}; \theta) = \exp\Big\{\sum_{s=1}^{n} \theta_s x_s + \sum_{s<t} \theta_{st} x_s x_t + \sum_{s<t<u} \theta_{stu} x_s x_t x_u + \ldots \\ \ldots + \theta_{1 \cdots N} x_1 x_2 \cdots x_N - \Phi(\theta)\Big\} \end{aligned} \tag{2.11}$$

It can be verified that the set of functions $\{x_s\}_{s=1}^N \cup \{x_s x_t\}_{s<t} \cup \ldots \cup \{x_1 \cdots x_n\}$ are linearly dependent over $\{0,1\}^N$, and span the space of all real-valued functions on $\{0,1\}^N$. Hence the dimension of the family is given by:

$$d(\theta) = \binom{N}{1} + \binom{N}{2} + \ldots + \binom{N}{N} = 2^N - 1$$

Since any distribution on the binary vector $\mathbf{x} \in \{0,1\}^N$ has $2^N - 1$ degrees of freedom,[5] we see that any distribution can be represented in the form equation (2.11).

Of course, the Ising model of equation (2.10) arises as a particular case of equation (2.11), where we place the restriction that $\theta_J = 0$ for all subsets $J \subset \{1, \ldots, N\}$ of size $|J| > 2$. Indeed, a nested sequence of exponential families $\mathcal{F}_k$ can be defined by imposing restrictions of the form $\theta_J = 0$ for all $|J| > k$, for $k = 1, \ldots, N-1$. See Amari [6] for details on such nested families. In the context of graphical models, these restrictions correspond to a limit on the maximal clique size in the associated graph.

Examples 2.2.2 and 2.2.3 can be extended to minimal exponential representations of m-ary processes ($m > 2$) as well. In particular, the analog of the Ising model for an m-ary process is specified in terms of the functions

$$\mathcal{R}(s) \triangleq \{x_s^a \mid a = 1, \ldots, m-1\} \qquad \text{for } s \in V \tag{2.12a}$$
$$\mathcal{R}(s,t) \triangleq \{x_s^a x_t^b \mid a, b = 1, \ldots, m-1\} \qquad \text{for } (s,t) \in E \tag{2.12b}$$

The dimension of this exponential family is given by $d(\theta) = (m-1)\,N + (m-1)^2|E|$. Incorporating higher order cliques entails adding higher degree multinomials to the clique functions of equation (2.12). This procedure, though conceptually straightforward, can lead to cumbersome notation. See Amari [6] for further details.

Overcomplete representations

In addition to such a minimal parameterization, parts of our analysis (especially Chapter 5) make use of an *overcomplete representation*, in which the $\{\phi_\alpha\}$ are linearly dependent. In this case, the lack of linear independence means that there exists an entire linear subspace of parameter vectors θ, each associated with the same distribution.

Example 2.2.4 (Overcomplete representation of binary process). One natural overcomplete representation of a binary process on a graph with pairwise cliques entails specifying a 2-vector for each node $s \in V$, and a 2×2 matrix of values for each edge (s,t) in the graph. To do so, we choose our clique potentials as indicator functions: that is, the collection of functions $\{\delta(x_s = j) \mid j = 0, 1\}$ for each node $s \in V$, and $\{\delta(x_s = j)\delta(x_t = k) \mid j, k = 0, 1\}$ for each edge $(s,t) \in E$. Here, the indicator or delta

[5] Any distribution can be represented by a 2^N vector, and we lose one degree of freedom due to the normalization $\sum_{\mathbf{x}} p(\mathbf{x}) = 1$.

function $\delta(x_s = j)$ is equal to 1 when node x_s takes the state value j, and 0 otherwise. The corresponding representation would be of the form

$$p(\mathbf{x};\theta) = \exp\big\{\sum_{s\in V}\sum_{j=0}^{1}\theta_{s;j}\delta(x_s = j) + \sum_{(s,t)\in E}\sum_{j,k=0}^{1}\theta_{st;jk}\delta(x_s = j)\delta(x_t = k) - \Phi(\theta)\big\} \tag{2.13}$$

where θ is the corresponding overcomplete parameter vector.

It is straightforward to generalize this type of overcomplete representation in terms of indicator functions to m-ary processes.

Different types of binary potentials

Given a distribution over a binary vector defined by a graph with pairwise cliques, it will often be useful to specify potential types from one of the following classes:

(a) in a graph with *attractive* potentials, all pairs of neighboring random variables are more likely to take the same values than opposite values.

(b) conversely, in a graph with *repulsive* potentials, all neighboring variables are encouraged to take opposite values.

(c) a graph with *mixed or frustrated* potentials consists of a combination of attractive and repulsive potentials.

In the statistical physics literature [e.g., 15, 32], these types of distributions are referred to as ferromagnetic, anti-ferromagnetic, and paramagnetic respectively.

The convention of this thesis will be that a binary random variable x_s takes values in $\{0,1\}$. In order to specify potential types, it is useful to consider a so-called *spin representation* in which a binary random variable u_s takes values in $\{-1,+1\}$. The term "spin" comes from the statistical physics literature [32]; for instance, one can think of u_s as giving the orientation (up or down) of a magnet at node s. We let

$$p(\mathbf{u};\omega) = \exp\{\sum_{s}\omega_s u_s + \sum_{(s,t)}\omega_{st}u_s u_t - \Phi(\omega)\} \tag{2.14}$$

be a minimal exponential representation corresponding to the spin vector $\mathbf{u} \in \{-1,1\}^N$, where ω is the associated vector of exponential parameters. In this spin representation, the nature of the interaction between u_s and u_t is determined entirely by the sign of ω_{st}. In particular, the potential is attractive (respectively repulsive) if and only if ω_{st} is positive (respectively negative).[6]

[6]Note that the same statement does not hold for the exponential parameter θ_{st} in a $\{0,1\}$ representation (see, e.g., equation (2.10)). For this representation, if we disregard the single node parameters θ_s, setting $\theta_{st} > 0$ places higher weight on the configuration $(x_s, x_t) = (1,1)$, but equal weights on the remaining configurations $\{(0,0),\ (1,0),\ (0,1)\ \}$.

Thus, a spin representation is convenient for constructing random distributions on graphs with particular types of potentials. Moreover, any spin parameter ω specifies a unique exponential parameter θ. In particular, we substitute the relation $u_s = 2x_s - 1$, which converts from $\{0,1\}$ variables to spins, into equation (2.14), and then equate coefficients with the Ising model of equation (2.10). In this way, we obtain the following relations:

$$\theta_s = 2\left[\omega_s - \sum_{t \in \mathcal{N}(s)} \omega_{st}\right] \tag{2.15a}$$

$$\theta_{st} = 4\omega_{st} \tag{2.15b}$$

We now define for future reference a few ensembles of random potentials. In all cases, we set the node parameters $\omega_s = 0$ for all nodes $s \in V$. Let $\mathcal{U}[a,b]$ denote the uniform distribution on the interval $[a,b]$. Given a particular *edge weight* $d > 0$, we then choose the edge parameters as follows:

(a) for the *uniform attractive ensemble* with edge weight $d > 0$, set $\omega_{st} \sim \mathcal{U}[0,d]$ independently for each edge $(s,t) \in E$

(b) for the *uniform repulsive ensemble*, set $\omega_{st} \sim \mathcal{U}[-d,0]$ independently for each edge $(s,t) \in E$

(c) for the *uniform mixed ensemble*, set $\omega_{st} \sim \mathcal{U}[-d,d]$ independently for each edge $(s,t) \in E$

Given the (randomly-chosen) distribution $p(\mathbf{u};\omega)$ specified in terms of the spin parameter ω, we then convert to the distribution $p(\mathbf{x};\theta)$, where θ is obtained from ω via equation (2.15).

■ 2.2.2 Properties of Φ

In this section, we develop some important properties of the log partition function Φ defined in equation (2.8b), including its convexity. Given a distribution $p(\mathbf{x};\theta)$ and a function $f : \mathcal{X}^N \rightarrow \mathbb{R}$, we define the expectation of $f(\mathbf{x})$ with respect to $p(\mathbf{x};\theta)$ as follows:

$$\mathbb{E}_\theta[f(\mathbf{x})] = \sum_{\mathbf{x} \in \mathcal{X}^N} p(\mathbf{x};\theta)\, f(\mathbf{x}) \tag{2.16}$$

When the sample space $\mathcal{X}$ is continuous, this summation should be replaced by an integral.

With this notation, we can show that the function Φ is closely related to the cumulant generating function[7] associated with the random variables $\{\phi_\alpha(\mathbf{x})\}$. In particular,

[7]Another interpretation of Φ arises in statistical physics, where it is known as the *Helmholtz free energy* [32, 136].

given a parameter vector $\theta \in \mathbb{R}^{d(\theta)}$ and another vector $\epsilon \in \mathbb{R}^{d(\theta)}$, we compute:

$$\log\Big(\mathbb{E}_\theta\big[\exp\{\sum_\alpha \epsilon_\alpha \phi_\alpha(\mathbf{x})\}\big]\Big) = \Phi(\theta+\epsilon) - \Phi(\theta) \tag{2.17}$$

The quantity on the left-hand side is the cumulant generating function (or the logarithm of the moment generating function) [81]. Equation (2.17) shows that this cumulant generating function is equal to the difference between the function Φ at two distinct values.

Using this relation, it can be shown that derivatives of Φ with respect to θ correspond to the cumulants of $\{\phi_\alpha(\mathbf{x})\}$. For example,

$$\frac{\partial \Phi}{\partial \theta_\alpha}(\theta) = \mathbb{E}_\theta[\phi_\alpha] \tag{2.18a}$$

$$\frac{\partial^2 \Phi}{\partial \theta_\alpha \, \partial \theta_\beta}(\theta) = \text{cov}_\theta\{\phi_\alpha, \, \phi_\beta\} \triangleq \mathbb{E}_\theta\Big\{\big(\phi_\alpha - \mathbb{E}_\theta[\phi_\alpha]\big)\big(\phi_\beta - \mathbb{E}_\theta[\phi_\beta]\big)\Big\} \tag{2.18b}$$

are the first and second order cumulants. In general, let $\text{cum}_\theta\{\phi_{\alpha_1}, \ldots, \phi_{\alpha_k}\}$ denote the k^{th}-order cumulant of $\{\phi_{\alpha_1}, \ldots, \phi_{\alpha_k}\}$ under $p(\mathbf{x};\theta)$. Then higher order cumulants are defined recursively by successive differentiation of lower order cumulants; e.g., $\text{cum}_\theta\{\phi_{\alpha_1}, \phi_{\alpha_2}, \phi_{\alpha_3}\} = \frac{\partial}{\partial \theta_{\alpha_3}}\big[\text{cum}_\theta\{\phi_{\alpha_1}, \phi_{\alpha_2}\}\big]$.

The second order cumulant in equation (2.18b) reveals an important property of the log partition function:

Lemma 2.2.1. *The function Φ is convex as a function of θ. The convexity is strict when the representation is minimal.*

Proof. Note that the quantity in (2.18b) is an element of the Fisher information matrix $(-\mathbb{E}_\theta\{\frac{\partial^2 \log p(\mathbf{x};\theta)}{\partial \theta^2}\})$. Therefore, the Hessian $\nabla^2\Phi$ is positive semi-definite (strictly positive definite for a minimal representation), so that Φ is convex (respectively strictly convex). □

The convexity of Φ will play a central role in subsequent geometric developments.

■ 2.2.3 Riemannian geometry of exponential families

An important feature of exponential families of distributions is their geometric structure. In this section, we provide a very brief introduction to the differential geometry of these families. See Amari [5–7] and the edited collection of papers [83] for further details.

Consider an exponential representation in terms of the $d(\theta)$-dimensional parameter θ, assumed to be minimal. For each $\theta \in \Theta$, we have $p(\mathbf{x};\theta) > 0$ for all $\mathbf{x} \in \mathcal{X}^N$. Therefore, we can associate with each point $\theta \in \Theta$ a function — namely, the log distribution $\log p(\mathbf{x};\theta)$. Under suitable regularity conditions [13], this association defines a $d(\theta)$-dimensional differential manifold $\mathcal{M}$ of functions $\{\log p(\mathbf{x};\theta) \mid \theta \in \Theta\}$. When the sample space $\mathcal{X}^N$ is discrete and hence finite, we can view $\mathcal{M}$ as embedded within $\mathbb{R}^{|\mathcal{X}^N|}$; otherwise, it is embedded within an infinite-dimensional function space. The mapping $\theta \mapsto \log p(\mathbf{x};\theta)$ is the co-ordinate mapping of the manifold, as illustrated in Figure 2.9.

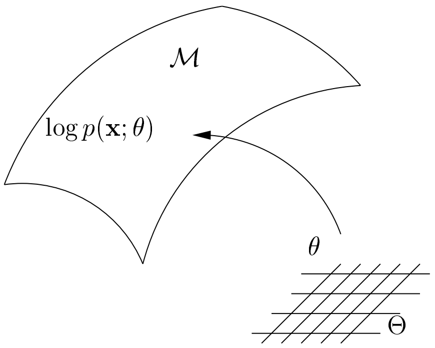


Figure 2.9. The exponential parameters θ serve as the co-ordinates for the $d(\theta)$-dimensional differential manifold of log distributions $\log p(\mathbf{x};\theta)$. Associated with each $\theta \in \Theta$ is a log distribution $\log p(\mathbf{x};\theta)$; the association $\theta \mapsto \log p(\mathbf{x};\theta)$ defines the co-ordinate mapping.

Given a line $\theta(t)$ in Θ, we can consider the curve in $\mathcal{M}$ defined by its image $\log p(\mathbf{x};\theta(t))$ under the co-ordinate mapping. The set of all tangent vectors to such curves at a particular value of θ defines the *tangent space* of $\mathcal{M}$ at the point $\log p(\mathbf{x};\theta)$. It can be seen that this tangent space is a $d(\theta)$-dimensional vector space. In particular, letting $\mathbf{e}_\alpha$ be a $d(\theta)$-vector of zeros with a single one in element α and zero elsewhere, consider the co-ordinate line $\theta(s;\alpha) = (1-s)\theta + s\mathbf{e}_\alpha$. By straightforward calculations, the tangent vector $\mathbf{t}_\alpha$ to the curve $\log p(\mathbf{x};\theta(s;\alpha))$ is given by

$$\mathbf{t}_\alpha = \frac{\partial}{\partial\theta_\alpha}\log p(\mathbf{x};\theta) \;=\; \phi_\alpha(\mathbf{x}) - \mathbb{E}_\theta\left[\phi_\alpha\right]\mathbf{1}(\mathbf{x}) \tag{2.19}$$

where $\mathbf{1}(\mathbf{x}) = 1$ for all $\mathbf{x} \in \mathcal{X}^N$. In computing this derivative, we have used equation (2.18a). It can be shown that the set $\{\mathbf{t}_\alpha \mid \alpha \in \mathcal{A}\}$ spans the tangent space at $\log p(\mathbf{x};\theta)$.

We now use $p(\mathbf{x};\theta)$ to define a weighted inner product on the tangent space. Of course, it suffices to specify the inner product for any pair $\{\mathbf{t}_\alpha, \mathbf{t}_\beta\}$, which we do as follows:

$$\langle \mathbf{t}_\alpha,\, \mathbf{t}_\beta\rangle_\theta = \mathbb{E}_\theta\Big[\frac{\partial}{\partial\theta_\alpha}\log p(\mathbf{x};\theta)\frac{\partial}{\partial\theta_\beta}\log p(\mathbf{x};\theta)\Big] = \mathrm{cov}_\theta\{\phi_\alpha,\, \phi_\beta\} \tag{2.20}$$

where we have used equations (2.18b) and (2.19) to see the equivalence to a covariance.

The quantities $g_{\alpha\beta}(\theta) \triangleq \text{cov}_\theta\{\phi_\alpha, \phi_\beta\}$ are elements of the *Fisher information matrix*, denoted by $G(\theta)$. For a minimal θ-representation, it can be seen that the Fisher information matrix is strictly positive definite for all θ. It therefore defines a Riemannian metric, with the squared distance between the distribution θ and an infinitesimally perturbed distribution $\theta + \Delta$ given by

$$[d(\theta, \theta + \Delta)]^2 \;=\; \sum_{\alpha,\beta} g_{\alpha\beta}(\theta)\Delta_\alpha \Delta_\beta \;=\; \Delta^T G(\theta)\Delta \;=\; \|\Delta\|^2_{G(\theta)} \tag{2.21}$$

The Fisher information matrix and the induced distance function of equation (2.21) also play important roles in other contexts, as we will explore in subsequent sections.

■ 2.2.4 Legendre transform and dual variables

The aspect of information geometry that sets it apart from classical Riemannian geometry is the existence of a dual parameterization, coupled to the exponential θ-parameterization. The coupling arises from convex duality associated with the log partition function Φ. The monograph of Rockafellar [151] provides a comprehensive treatment of convex duality; a more elementary and geometric treatment of duality can be found in Bertsekas [20]. In this section, we exploit the convexity of Φ to apply notions from convex analysis — in particular, the Legendre transform — from which we obtain a second set of parameters dual to the exponential θ-representation. In a later section, we use the Legendre duality to develop a geometric interpretation of the presence or absence of certain cliques in a graph-structured distribution.

The convexity of Φ allows us to apply the Legendre transform. Here we assume that the domain Θ of θ is either all of $\mathbb{R}^{|\mathcal{A}|}$, or some convex subset. The Legendre dual of Φ is defined as:

$$\Psi(\eta) = \sup_\theta\{\eta^T\theta - \Phi(\theta)\} \tag{2.22}$$

where η is a vector of the same dimension as the exponential parameter θ. Since the quantity to be maximized (i.e., $\eta^T\theta - \Phi(\theta)$) is strictly concave as a function of θ, the supremum in equation (2.22) is attained at some point $\widehat{\theta}$. Taking derivatives to find stationary points, and making use of equation (2.18a) yields the defining equation:

$$\eta_\alpha \;\equiv\; \eta_\alpha(\widehat{\theta}) \;=\; \mathbb{E}_{\widehat{\theta}}[\phi_\alpha] \tag{2.23}$$

Since they are obtained by taking expectations, these dual variables η are often referred to as the *mean parameters*. Substituting the relation in (2.23) back into equation (2.22) yields the relation

$$\Psi(\eta(\widehat{\theta})) \;=\; \sum_\alpha \widehat{\theta}_\alpha \mathbb{E}_{\widehat{\theta}}[\phi_\alpha] - \Phi(\widehat{\theta}) \;=\; \mathbb{E}_{\widehat{\theta}}[\log p(\mathbf{x}; \widehat{\theta})] \tag{2.24}$$

so that the Legendre dual Ψ is the negative entropy. Note that Ψ is itself a convex function, so that we can again apply the Legendre transform. It is not difficult to show that applying the Legendre transform twice in this manner recovers the log partition function; that is,

$$\Phi(\theta) = \sup_{\eta}\{\theta^T\eta - \Psi(\eta)\} \tag{2.25}$$

The Legendre duality of Φ gives rise to a mapping $\Lambda : \theta \mapsto \eta$, defined explicitly by

$$[\Lambda(\theta)]_\alpha = \frac{\partial\Phi(\theta)}{\partial\theta_\alpha} = \mathbb{E}_\theta[\phi_\alpha] \tag{2.26}$$

For a minimal representation, Lemma 2.2.1 guarantees that Φ is strictly convex, in which case the mapping is one-to-one [151]. It is therefore invertible on its image, with the inverse map $\Lambda^{-1} : \eta \mapsto \theta$ defined by the corresponding relation

$$[\Lambda^{-1}(\eta)]_\alpha = \frac{\partial\Psi(\eta)}{\partial\eta_\alpha} \tag{2.27}$$

On the basis of these mappings, we can specify distributions either in terms of the exponential parameter θ, or the associated dual parameter η. Given a valid dual parameter η in a minimal representation, the quantity $p(\mathbf{x};\eta)$ denotes the equivalent exponential distribution $p(\mathbf{x};\Lambda^{-1}(\eta))$.

A few examples help to give intuition for the Legendre mapping:

Example 2.2.5 (Legendre transform for Gaussian). Let $\mathbf{x} \sim \mathcal{N}(0, P)$ be a zero-mean Gaussian random vector with covariance P. Then the density has an exponential representation of the form:

$$p(\mathbf{x};\theta) = \exp\Big\{\frac{1}{2}\sum_{s=1}^{N}\theta_{ss}x_s^2 + \sum_{s<t}\theta_{st}x_s x_t - \Phi(\theta)\Big\} \tag{2.28}$$

Here θ specifies elements of the inverse covariance (i.e., $P^{-1}_{st} = -\theta_{st}$).[8] From equation (2.26), the dual variables are given by:

$$\begin{aligned}\eta_{ss} &= \mathbb{E}_\theta[x_s^2] = \mathrm{var}(x_s)\\ \eta_{st} &= \mathbb{E}_\theta[x_s x_t] = \mathrm{cov}(x_s, x_t)\end{aligned}$$

so that η specifies elements of the covariance matrix P. That is, the Legendre transform maps back and forth between the matrix inverse pair P and $-P^{-1}$.

Example 2.2.6. We now return to the Ising model (see Example 2.2.2), where the random vector $\mathbf{x} \in \{0,1\}^N$ has a distribution of the form

$$p(\mathbf{x};\theta) = \exp\{\sum_{s\in V}\theta_s x_s + \sum_{(s,t)\in E}\theta_{st}x_s x_t - \Phi(\theta)\}$$

[8] We require that θ belongs to the set for which $P^{-1}(\theta) > 0$.

In this case, the dual variables are given by the expectations:

$$\begin{aligned} \eta_s &= \mathbb{E}_\theta[x_s] \equiv p(x_s = 1; \theta) \\ \eta_{st} &= \mathbb{E}_\theta[x_s x_t] \equiv p(x_s = 1, x_t = 1; \theta) \end{aligned}$$

That is, the dual variables correspond to particular marginal probabilities at individual nodes, and pairs of nodes (s,t). Note that the dual variables fully specify the single node marginals, and pairwise marginals for $(s,t) \in E$:

$$p(x_s; \theta) = [(1-\eta_s) \;\; \eta_s]^T \; ; \qquad p(x_s, x_t; \theta) = \begin{bmatrix} (1 + \eta_{st} - \eta_s - \eta_t) & \eta_t - \eta_{st} \\ \eta_s - \eta_{st} & \eta_{st} \end{bmatrix}$$

When the underlying graph G of the Ising model is a tree, these local marginals determine the full distribution $p(\mathbf{x}; \theta)$ explicitly via the tree factorization given in equation (2.6). For a graph with cycles, such a local construction is not possible; however, whenever the dual variables $\{\eta_s, \eta_{st}\}$ belong to the range of the Legendre transform, then the invertibility of this transform guarantees that the dual variables still completely specify the distribution.

The Legendre mapping is also closely related to the Fisher information matrix. In particular, by differentiating equation (2.23) with respect to θ_β, we see that the Jacobian of the mapping $\Lambda : \theta \mapsto \eta$ is given by the Fisher information matrix $[G(\theta)]_{\alpha\beta} = \mathrm{cov}_\theta\{\phi_\alpha, \, \phi_\beta\}$. That is,

$$\eta(\theta + \Delta\theta) - \eta(\theta) \approx G(\theta)\Delta\theta \tag{2.29}$$

up to first order in the perturbation $\Delta\theta$. Similarly, the inverse Fisher information matrix G^{-1}, which is guaranteed to exist when Φ is strictly convex, corresponds to the Jacobian of the inverse mapping $\Lambda^{-1} : \eta \mapsto \theta$.

■ 2.2.5 Geometric consequences for graphical models

In the specific context of graphical models, the Legendre duality also leads to an interesting geometric interpretation of the presence or absence of given clique potentials. In particular, consider the constrained maximum entropy problem:

$$\begin{cases} \max_p H(p) \\ \sum_{\mathbf{x}} p(\mathbf{x}) \phi_\alpha(\mathbf{x}) \leq \eta_\alpha \end{cases} \tag{2.30}$$

Geometrically, the maximization takes place over a polyhedral set, formed by the intersection of the probability simplex $\mathcal{P} = \{p(\mathbf{x}) \mid 0 \leq p(\mathbf{x}) \leq 1; \; \sum_{\mathbf{x}} p(\mathbf{x}) = 1 \}$ with the hyperplane constraints $\{p(\mathbf{x}) \mid \sum_{\mathbf{x}} p(\mathbf{x}) \phi_\alpha(\mathbf{x}) \leq \eta_\alpha \}$. It is well-known [42] that the solution to this problem assumes the familiar Gibbs form of equation (2.8), where the exponential parameter θ_α now corresponds to the Lagrange multiplier associated with the constraint $\sum_{\mathbf{x}} p(\mathbf{x})\phi_\alpha(\mathbf{x}) \leq \eta_\alpha$. That is, it reflects the sensitivity of the problem to perturbations in the associated η_α constraint.

By the Karush-Kuhn-Tucker conditions [20], the Lagrange multiplier θ_α is zero whenever the η_α-constraint is inactive (i.e., not met with equality.) On this basis, the presence or absence of particular cliques in a graphical model can be related to hyperplane constraints. In particular, we can add a given clique potential ϕ_β by imposing a hyperplane constraint of the form $\sum_{\mathbf{x}} p(\mathbf{x})\phi_\beta(\mathbf{x}) \leq \eta_\beta$. Progressively lowering η_β so as to tighten the constraint will eventually ensure that the associated Lagrange multiplier is non-zero, meaning that the clique potential ϕ_β appears in the exponential representation with a non-zero weight θ_β. Conversely, we can remove a given clique from the graphical distribution by loosening the associated constraint. Eventually, the constraint will become inactive, so that the Lagrange multiplier θ_β is zero and the clique is effectively absent from the graph.

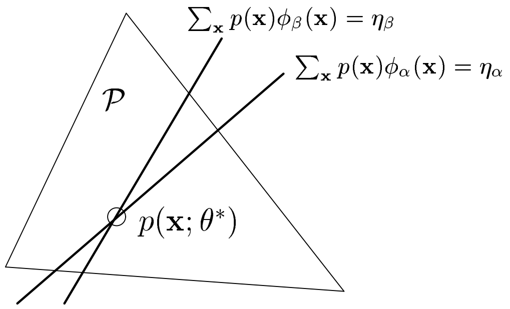


Figure 2.10. Geometry of graph-structured distributions. Distributions $p(\mathbf{x})$ are restricted to the simplex $\mathcal{P} = \{\ p(\mathbf{x}) \mid 0 \leq p(\mathbf{x}) \leq 1;\ \ \sum_{\mathbf{x}} p(\mathbf{x}) = 1\ \}$, and lie on the intersections of hyperplane constraint sets $\{\ p(\mathbf{x}) \mid \sum_{\mathbf{x}} p(\mathbf{x})\phi_\alpha(\mathbf{x}) = \eta_\alpha\ \}$ imposed by the clique potentials $\{\phi_\alpha\}$.

■ 2.2.6 Kullback-Leibler divergence and Fisher information

The Kullback-Leibler divergence [119] can be viewed as a measure of the "distance" between two distributions. For a discrete random vector, its usual definition [see, e.g., 42] is $D(p \parallel q) = \sum_{\mathbf{x}\in\mathcal{X}^N} p(\mathbf{x})[\log p(\mathbf{x}) - \log q(\mathbf{x})]$. The definition shows that it is not a true distance, since (for example) it is not symmetric in p and q. However, it can be shown using Jensen's inequality that $D(p \parallel q) \geq 0$ for all p and q, with equality if and only if $p = q$.

With a minor abuse of notation[9], we let $D(\theta \parallel \theta^*)$ denote the KL divergence between two distributions in exponential form $p(\mathbf{x};\theta)$ and $p(\mathbf{x};\theta^*)$. The exponential parameter-

[9]Strictly speaking, the divergence applies to distributions $p(\mathbf{x};\theta)$ and $p(\mathbf{x};\theta^*)$, and not to the parameters θ and θ^* themselves.

ization leads to an alternative representation of this KL divergence:

$$\begin{aligned} D(\theta \parallel \theta^*) &= \sum_\alpha \mathbb{E}_\theta[\phi_\alpha]\big[\theta - \theta^*\big]_\alpha + \Phi(\theta^*) - \Phi(\theta) \\ &= \eta^T[\theta - \theta^*] + \Phi(\theta^*) - \Phi(\theta) \end{aligned} \tag{2.31}$$

where $\eta_\alpha = [\Lambda(\theta)]_\alpha$. That is, the pair (θ, η) are dually coupled via the Legendre transform.

Equation (2.31) shows that the KL divergence $D(\theta \parallel \theta^*)$ can be viewed as a Bregman distance, induced by the convex log partition function Φ. In particular, since $\frac{\partial \Phi(\theta)}{\partial \theta_\alpha} = \mathbb{E}_\theta[\phi_\alpha]$, the KL divergence $D(\theta \parallel \theta^*)$ is equivalent to the difference between $\Phi(\theta^*)$ and the first-order tangent approximation $\Phi(\theta) + \nabla^T \Phi(\theta)(\theta^* - \theta)$, as illustrated

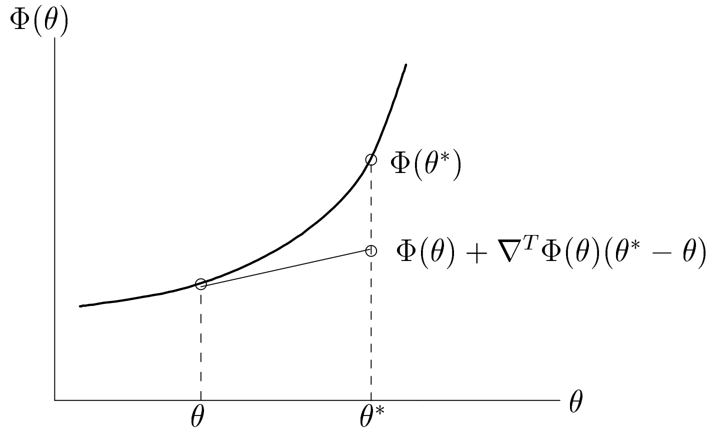


Figure 2.11. Kullback-Leibler divergence as a Bregman distance induced by the log partition function Φ. The KL divergence $D(\theta \parallel \theta^*)$ is equal to the difference between $\Phi(\theta^*)$ and the tangent approximation $\Phi(\theta) + \nabla^T \Phi(\theta)(\theta^* - \theta)$.

in Figure 2.11. Bregman distances are defined in precisely this manner; see Censor and Zenios [31] for more details on Bregman distances and their properties. For a minimal representation, the strict convexity of Φ guarantees that this tangent approximation is always an underestimate of $\Phi(\theta^*)$, so that the KL divergence is positive for $\theta \neq \theta^*$.[10]

It is also possible to re-write the KL divergence in terms of the dual variables η. In particular, from the Legendre duality between Φ and Ψ, we have for all dually coupled pairs (θ, η):

$$\eta^T \theta = \Phi(\theta) + \Psi(\eta) \tag{2.32}$$

Substituting this relation into equation (2.31), we obtain an alternative representation

[10] For overcomplete representations, it is possible to have distinct parameters $\theta \neq \theta^*$ that induce the same distribution, in which case $D(\theta \parallel \theta^*) = 0$.

of the KL divergence:

$$D(\theta \parallel \theta^*) = (\theta^*)^T[\eta^* - \eta] + \Psi(\eta) - \Psi(\eta^*) \tag{2.33}$$

Comparing equations (2.33) and (2.31), we see that the former is obtained from the latter by replacing Φ with its dual Ψ, and interchanging the roles of θ and θ^* (and their associated dual parameters η and η^*). Equation (2.33) gives rise to the notion of the dual of the KL divergence, as studied by Chentsov [33, 34].

The Kullback-Leibler divergence is very closely related to the Riemannian metric defined in equation (2.21). In particular, by Taylor series expansion of $\log p(\mathbf{x}; \theta)$, we obtain

$$D(\theta \parallel \theta^*) \approx \frac{1}{2}[\theta - \theta^*]^T G(\theta)[\theta - \theta^*] \;=\; \frac{1}{2}\|\theta - \theta^*\|^2_{G(\theta)} \tag{2.34}$$

where the approximate equality holds up to second order. In this sense, the squared distance induced by the Fisher information $G(\theta)$ is an approximation to the KL divergence. This notion will arise again in Section 2.2.8.

2.2.7 I-projections onto flat manifolds

In this section, we define a pair of optimization problems canonical to information geometry. In particular, they entail projecting (where the KL divergence serves as the "distance") a given distribution onto certain types of "flat" manifolds. The dual parameterizations allow us to specify two types of flat manifold, depending on whether distributions are specified in terms of the exponential parameters θ, or the mean parameters η. This procedure of projecting onto a flat manifold, known as an *I-projection*, constitutes the basic building block for a variety of well-known optimization algorithms [e.g., 44, 135].

We begin with definitions of e and m-flat manifolds:

Definition 2.2.1. Given a linear subset of Θ, an *e-flat manifold* corresponds to its image under the coordinate mapping $\theta \mapsto p(\mathbf{x}; \theta)$. That is,

$$\mathcal{M}_e = \{\ p(\mathbf{x}; \theta) \mid A\theta = a\ \} \tag{2.35}$$

for some matrix A and vector a.

An *e-geodesic* is a 1-dimensional e-flat manifold — that is, a family of distributions specified by a line in the exponential coordinates:

$$\{\ p(\mathbf{x}; \theta(t)) \;\mid\; \theta(t) = (1-t)\,\theta_0 + t\,\theta_1,\;\; t \in \mathbb{R}\ \}$$

for some fixed θ_0 and θ_1.

With a minor abuse of notation, we shall often use $\theta \in \mathcal{M}_e$ to mean that θ belongs to the linear subset defining the e-flat manifold. To illustrate, we consider a few examples:

Example 2.2.7. For the Ising model (Example 2.2.2), an important type of e-flat manifold is induced by the linear subset $\mathcal{F}_0 = \{ \theta \mid \theta_{st} = 0 \;\; \forall \;\; (s,t) \;\in\; E \}$. Any distribution in this family has the representation

$$p(\mathbf{x};\theta) = \exp\big\{ \sum_{s=1}^{N} \theta_s x_s - \Phi(\theta) \big\}$$

That is, these distributions are fully factorized, with no interactions between different components x_s of the random vector $\mathbf{x}$.

We specify an m-flat manifold in a similar fashion:

Definition 2.2.2. An *m-flat manifold* is the set of distributions corresponding to a linear subset of the dual variables:

$$\mathcal{M}_m = \{ p(\mathbf{x};\; \eta) \mid B\eta = b \} \tag{2.36}$$

Recall that $p(\mathbf{x};\eta)$ refers to the exponential distribution given by $p(\mathbf{x};\Lambda^{-1}(\eta))$.

We define an m-geodesic in a similar fashion to an e-geodesic: that is, as a 1-dimensional family of distributions specified by a line in the dual coordinates. Again, we shall often abuse notation by writing $\eta \in \mathcal{M}_m$ to mean that η belongs to the linear subset defining $\mathcal{M}_m$.

Note that an m-geodesic corresponds to the familiar type of mixture of distributions. That is, given the line $\eta(t) = (1-t)\eta_0 + t\eta_1$, the induced m-geodesic corresponds to mixtures of distributions in the form

$$p(\mathbf{x};\eta(t)) = (1-t)\, p(\mathbf{x};\eta_0) + t\, p(\mathbf{x};\eta_1)$$

We consider a few examples of m-flat manifolds:

Example 2.2.8. Consider a scalar Gaussian random variable with exponential representation $p(x;\theta) = \exp\{\theta_1 x + \theta_2 x^2 - \Phi(\theta)\}$. Here the dual parameters are given by $\eta_1 = \mathbb{E}_\theta[x] = \mu$, and $\eta_2 = \mathbb{E}_\theta[x^2] = \mu^2 + \sigma^2$, where μ and σ are the mean and standard deviation respectively. Thus, we see that the set of scalar Gaussian distributions with fixed mean corresponds to an m-flat manifold.

Example 2.2.9. Consider the Ising model (see Example 2.2.2). The mean parameters consist of the probabilities $\eta_s = p(x_s = 1;\theta)$ and $\eta_{st} = p(x_s = 1, x_t = 1;\theta)$. Thus, the set of all distributions with a specified set of single-node marginals

$$\{ \; p(\mathbf{x};\eta) \mid \eta_s = \widetilde{\eta}_s \; \}$$

forms an m-flat manifold.

The notions of e-flat and m-flat manifolds give rise to a pair of canonical optimization problems in information geometry. We begin by considering the problem of projecting onto an e-flat manifold.

Projection onto an e-flat manifold

For a fixed reference distribution θ^* and e-flat manifold $\mathcal{M}_e$, consider the constrained optimization problem:

$$\begin{cases} \min_\theta D(\theta^* \parallel \theta) \\ \text{s.t} \qquad \theta \in \mathcal{M}_e \end{cases} \tag{2.37}$$

From equation (2.31) and the convexity of Φ, we see that the KL divergence $D(\theta^* \parallel \theta)$ is a convex function of its second argument. Therefore, problem (2.37) is a convex optimization problem with linear constraints, so that it has a unique global optimum — say $\widehat{\theta} = \arg\min_{\theta \in \mathcal{M}_e} D(\theta^* \parallel \theta)$. Using equations (2.18a), and (2.31), we compute the gradient $\nabla_\theta D(\theta^* \parallel \theta) = \eta - \eta^*$. By the standard condition for a global minimum of a convex function over a linear manifold [20], we obtain:

$$[\eta^* - \widehat{\eta}]^T \, [\theta - \widehat{\theta}] = 0 \tag{2.38}$$

for all $\theta \in \mathcal{M}_e$. Equation (2.38) is the defining condition for $\widehat{\theta}$, which is known as the *I-projection* of the point θ^* onto $\mathcal{M}_e$.

Many e-flat manifolds of interest are obtained by zeroing a subset of the exponential parameters — that is:

$$\mathcal{F}_{\mathcal{J}} = \{\theta \mid \theta_\alpha = 0 \quad \forall\, \alpha \notin \mathcal{J}\,\}$$

The set of fully factorized distributions, described in Example 2.2.7, is an important case. The optimality condition of equation (2.38) has strong consequences for projections onto such manifolds. In particular, for any index $\beta \in \mathcal{J}$, we can form a perturbation $\Delta\theta = \mathbf{e}_\beta$ of all zeros except for a one in the β-entry. This perturbation $\Delta\theta$ lies in the e-flat manifold $\mathcal{F}_{\mathcal{J}}$, so that it must be orthogonal to $[\eta^* - \widehat{\eta}]$. Using equation (2.38), this implies that

$$\eta^*_\beta = \widehat{\eta}_\beta \qquad \forall\, \beta \,\in\, \mathcal{J} \tag{2.39}$$

That is, the dual parameters of the projection $\widehat{\eta}$ must agree with the dual parameters η^* of the original distribution for all indices β that are free to vary.

Example 2.2.10. Consider again the Ising model (Example 2.2.2), and consider the problem of projecting θ^* onto the set $\mathcal{F}_0 = \{\ \theta \mid \theta_{st} = 0 \ \ \forall \ \ (s,t) \ \in \ E\ \}$ of fully factorized distributions (see Example 2.2.7). Then equation (2.39) ensures that the I-projection $\widehat{\theta}$ satisfies:

$$\mathbb{E}_{\widehat{\theta}}[x_s] \ = \ \widehat{\eta}_s \ = \ \eta^*_s \ = \ \mathbb{E}_{\theta^*}[x_s]$$

Since $\mathbf{x} \in \{0,1\}^N$ is a binary random variable, the dual variables are equivalent to node marginal probabilities. Therefore, the single node marginals of the I-projection $p(\mathbf{x};\widehat{\theta})$ agree with those of $p(\mathbf{x};\theta^*)$. This type of property holds for more general nested families of distributions, as described in Amari [6].

Projection onto a m-flat manifold

The problem of projecting onto an m-flat manifold $\mathcal{M}_m$ is effectively dual to the problem of projecting onto an e-flat manifold. In this case, the relevant optimization problem is

$$\begin{cases} \min_\theta D(\theta \,\|\, \theta^*) \\ \text{s.t} \qquad \theta \,\in\, \mathcal{M}_m \end{cases} \tag{2.40}$$

Note that in contrast to problem (2.37), here we optimize over the *first* argument of the KL divergence. This change should be understandable, in light of the relation between the KL divergence in (2.31), and its alternative formulation in equation (2.33).

On the basis of equation (2.33), it can be shown that problem (2.40) is convex and linearly constrained in the dual variables η, and so has a unique global optimum $\widehat{\eta}$. Again, straightforward computations yield the defining condition for this optimum:

$$[\theta^* - \widehat{\theta}]^T \, [\eta - \widehat{\eta}] = 0 \qquad \forall\, \eta \,\in\, \mathcal{M}_m \tag{2.41}$$

2.2.8 Geometry of I-projection

Associated with equations (2.38) and (2.41) is an elegant geometric picture. Here we present this viewpoint for the I-projection onto an e-flat manifold, noting that a similar picture holds for I-projection onto an m-flat manifold. The Pythagorean results of this section (Theorems 2.2.1 and 2.2.2) date back to Kullback [119,120] and Csiszár [44,45].

To develop the geometry, note first of all that for any $\theta \in \mathcal{M}_e$, the vector $[\theta - \widehat{\theta}]$ can be viewed as the tangent vector to some e-geodesic lying within $\mathcal{M}_e$, as illustrated in Figure 2.12. Secondly, consider the m-geodesic joining the points θ^* and $\widehat{\theta}$. Although it is linear by definition in η-coordinates, it will be a curve in the θ-coordinates — namely:

$$\theta(t) = \Lambda^{-1}(\widehat{\eta} + t[\eta^* - \widehat{\eta}]) \tag{2.42}$$

This curved m-geodesic is illustrated in Figure 2.12. We calculate the tangent vector to the curve (2.42) at $\widehat{\eta}$ (i.e., at $t = 0$) as $G^{-1}(\widehat{\eta})\; [\eta^* - \widehat{\eta}]$, where we have recalled that the Jacobian of the inverse mapping Λ^{-1} is given by the inverse Fisher information G^{-1}.

Now consider the inner product, as defined by the Fisher information matrix $G(\widehat{\theta})$, between these two tangent vectors. In particular, using the fact that $G(\widehat{\theta})G^{-1}(\widehat{\eta}) = I$, we compute:

$$\langle [\theta - \widehat{\theta}],\, G^{-1}(\widehat{\eta})\; [\eta^* - \widehat{\eta}] \rangle_{G(\widehat{\theta})} \quad = \quad [\eta^* - \widehat{\eta}]^T \; [\theta - \widehat{\theta}] \tag{2.43}$$

which must vanish by equation (2.38). Therefore, the geometric consequence of equation (2.38) is that the m-geodesic joining θ^* and $\widehat{\theta}$ forms an orthogonal intersection with the e-flat manifold $\mathcal{M}_e$, as illustrated in Figure 2.12. Here orthogonality on the left side of equation (2.43) is measured using the Riemannian inner product $\langle\, , \,\rangle_{G(\widehat{\theta})}$ induced by the Fisher information matrix (see Section 2.2.3), whereas the right side corresponds to Euclidean inner product of two vectors in $\mathbb{R}^{d(\theta)}$. Since the I-projection

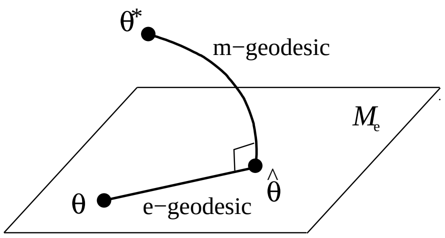


Figure 2.12. The point θ^* is projected onto the e-flat manifold $\mathcal{M}_e$ by following an m-geodesic. This yields the I-projection $\widehat{\theta}$. The tangent vector to the m-geodesic joining θ^* and $\widehat{\theta}$ is orthogonal to the manifold $\mathcal{M}_e$.

of θ^* onto an e-flat manifold is obtained by following an m-geodesic, it is often called an m-*projection.*

Recall from equation (2.34) that the KL divergence is closely related to the Riemannian metric induced by the Fisher information matrix. The geometric picture of I-projection allows us to further develop this relation by showing that, as with Hilbert space norms, the KL divergence satisfies a type of Pythagorean relation:

Theorem 2.2.1 (Pythagoras for m-projection). *Let $\widehat{\theta}$ be the I-projection of a point θ^* onto an e-flat manifold $\mathcal{M}_e$. Then for all $\theta \in \mathcal{M}_e$, we have:*

$$D(\theta^* \parallel \theta) = D(\theta^* \parallel \widehat{\theta}) + D(\widehat{\theta} \parallel \theta) \tag{2.44}$$

Proof. We provide the proof here, since it follows in a straightforward manner from our earlier development. We first use the form of the KL divergence in (2.31) to write:

$$D(\theta^* \parallel \widehat{\theta}) + D(\widehat{\theta} \parallel \theta) \quad = \quad [\eta^*]^T[\theta^* - \widehat{\theta}] + \widehat{\eta}^T[\widehat{\theta} - \theta] + \Phi(\theta^*) - \Phi(\theta)$$

We then use the optimality condition of equation (2.38) to rewrite the second term on the RHS as $[\eta^*]^T[\widehat{\theta} - \theta]$. Cancelling out the $[\eta^*]^T\widehat{\theta}$ terms then yields the result. □

Figure 2.12 again illustrates this geometry, where the points θ, $\widehat{\theta}$ and θ^* correspond to the vertices of a "right" triangle, with the segment between θ and θ^* corresponding to the hypotenuse. The "distances" between these three points are related by the Pythagorean relation[11] of equation (2.44).

We note that a geometric picture similar to that of Figure 2.12 also holds for the I-projection of θ^* (or alternatively, η^*) onto an m-flat manifold. The primary difference is that the picture holds in the dual coordinates η, rather than the exponential coordinates.

[11] In passing, we note that this Pythagorean relation holds more generally for projections onto linear sets, where the projection is defined by any Bregman distance. See [31] for further details on Bregman distances and their properties.

In this case, the projection $\widehat{\eta}$ is obtained by following a e-geodesic (curved in the η-coordinates) between η^* and $\widehat{\eta}$. For this reason, this operation is often called an e-*projection* (onto an m-flat manifold). Moreover, a Pythagorean relation analogous to that of equation (2.44) also holds:

Theorem 2.2.2 (Pythagoras for e-projection). *Let $\widehat{\theta}$ be the I-projection of a point θ^* onto an m-flat manifold $\mathcal{M}_m$. Then for all $\theta \in \mathcal{M}_m$, we have:*

$$D(\theta \parallel \theta^*) = D(\theta \parallel \widehat{\theta}) + D(\widehat{\theta} \parallel \theta^*)$$

Proof. The proof of this result is entirely analogous to that of Theorem 2.2.1. □

Various extensions to Theorems 2.2.1 and 2.2.2 are possible. For example, if we project onto a convex set of distributions (as opposed to a m-flat or linear manifold), then the equality of Theorem 2.2.2 is weakened to an inequality (i.e., from a Pythagorean result to the triangle inequality) [see 42].

Moreover, I-projections constitute the basic building blocks for a variety of well-known iterative algorithms. These algorithms can be divided into two broad classes: *successive projection algorithms*, and *alternating minimization algorithms*. Csiszár [45] established the convergence of the successive projection technique; in a later paper [47], he showed that the iterative scaling procedure [52] is a particular case of such an algorithm. Csiszár and Tusnády [44] present alternating minimization algorithms, and provide conditions for their convergence. The tutorial introduction by O'Sullivan [135] shows how many well-known algorithms (e.g., expectation-maximization [57], Blahut-Arimoto [9, 24]) can be reformulated as particular cases of alternating minimization.

■ 2.3 Variational methods and mean field

The term *variational methods* refers to a variety of optimization problems, and associated techniques for their solution. Its origins lie in the calculus of variations [73], where the basic problem is finding the extremum of an integral involving an unknown function and its derivatives. Modern variational methods encompass a wider range of techniques, including the finite element method [157], dynamic programming [19], as well as the maximum entropy formalism [99, 100, 182]. Here we begin with a simple example to motivate the idea of a variational method; we then turn to an exposition of mean field methods. For more details, we refer the reader to the tutorial paper [93], which provides an introduction to variational methods with emphasis on their application to graphical models. The book by Rustagi [153] gives more technical details, with particular applications to problems in classical statistics.

To motivate the idea of a variational method, consider the following example, also discussed in the tutorial paper [93]. Suppose that for a fixed vector $b \in \mathbb{R}^n$ and symmetric positive definite matrix $Q \in \mathbb{R}^{n \times n}$, we are interested in solving the linear equation $Qx = b$. Clearly, the solution is $\widehat{x} = Q^{-1}b$, which could be obtained by performing a brute force matrix inversion, and then forming a matrix-vector product. For large problems, this brute force approach will be intractable. A variational formulation of the problem motivates a more efficient technique, and suggests natural approximations to the optimum. In particular, we consider the cost function $J(x) = \frac{1}{2}x^T Qx - b^T x$. Clearly, $J(x)$ is convex and bounded below, and so has a unique global minimum. Indeed, the minimizing argument of $J(x)$ is the desired optimum; that is, we can compute $\widehat{x}$ by minimizing $J(x)$. Moreover, to obtain approximations to the optimum, we need only perform a partial minimization of $J(x)$. The method of choice for such problems is the conjugate gradient method of numerical linear algebra [56]. It generates a sequence $\{x^k\}$, such that each x^k minimizes $J(x)$ over a $k-$dimensional subspace. Thus, the n^{th} iterate x^n will be equal (aside from possible numerical inaccuracies) to the optimum $\widehat{x}$; however, the iterations are typically terminated for some $k \ll n$, thereby yielding an approximation $x^k \approx \widehat{x}$.

2.3.1 Mean field as a variational technique

We now describe particular subclass of variational methods known under the rubric of *mean field*. This term refers to a collection of techniques for obtaining approximations to distributions. While we take a variational approach to mean field, these methods can be motivated and derived from a variety of perspectives [e.g., 32, 136]. Our exposition shares the spirit of the tutorial introductions given in [93, 106]; it differs in details in that we make extensive use of exponential representation of equation (2.8).

Let $p(\mathbf{x};\theta^*)$ be the distribution of interest. We assume that this distribution is intractable, so approximating it is a natural problem. Consider the variational problem of minimizing the Kullback-Leibler divergence

$$D(\theta \parallel \theta^*) = \sum_{\alpha} \mathbb{E}_{\theta}[\phi_\alpha]\, [\theta - \theta^*]_\alpha + \Phi(\theta^*) - \Phi(\theta) \tag{2.45}$$

between $p(\mathbf{x};\theta)$ and $p(\mathbf{x};\theta^*)$. Of course, if we could perform an unconstrained minimization, we would trivially recover $\theta \equiv \theta^*$. However, since calculating the KL divergence in (2.45) entails taking expectations under $p(\mathbf{x};\theta)$, it is necessary to restrict the minimization to a tractable family $\mathcal{F}$ of distributions. In particular, we form an approximation $p(\mathbf{x};\widehat{\theta})$ by computing:

$$\widehat{\theta} = \arg\min_{\theta\in\mathcal{F}} D(\theta \parallel \theta^*) \tag{2.46}$$

That is, we compute the optimal approximation in some family, where optimality is measured by the KL divergence. It is important that this optimization problem does *not* correspond to an I-projection. Indeed, although we will see that $\mathcal{F}$ typically corresponds to an e-flat manifold, the optimization in equation (2.46) is over the first argument of the KL divergence, and not the second as it would be for a projection onto an e-flat manifold (see problem (2.37)). The fact that mean field is not an I-projection has important consequences, as we will discuss later.

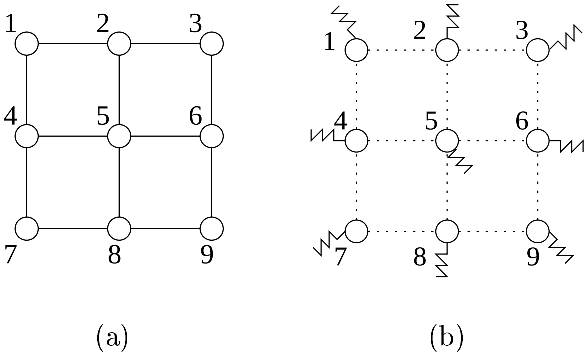


Figure 2.13. Illustration of the mean field approximation. (a) Original graph is a 3×3 grid. (b) Fully disconnected graph, corresponding to a naive mean field approximation. Wavy lines at each node represent adjustable input parameters.

The form of the KL divergence in (2.45) suggests an alternative interpretation of this optimization. By the convexity of the log partition function and equation (2.18a), we have

$$\Phi(\theta^*) \geq \Phi(\theta) + \sum_{\alpha} \mathbb{E}_{\theta}[\phi_\alpha] \, [\theta^* - \theta]_\alpha \tag{2.47}$$

for all θ. This lower bound also appears in the statistical physics literature, with slightly different notation, as the so-called Gibbs-Bogoliubov-Feynman inequality [see 32, 185]. As a consequence of equation (2.47), the optimization of (2.46) can be viewed as *maximizing* a lower bound on the (intractable) log partition function $\Phi(\theta^*)$. This interpretation is important in the application of mean field methods to parameter estimation via the EM algorithm [see 106].

The formulation in (2.46) encompasses a variety of mean field techniques, where a specific technique corresponds to a particular choice of e-flat manifold for the approximating family $\mathcal{F}$. For example, given the Ising model of equation (2.10), consider the family $\mathcal{F}_0 = \{\theta \mid \theta_{st} = 0 \;\; \forall \, (s,t) \; \in \; E\,\}$ — that is, the e-flat manifold of fully factorized distributions (see Example 2.2.7)). Performing the minimization of equation (2.46) with this choice corresponds to finding the best fully factorized approximation. Doing so entails finding zero points of the gradient, and a particular iterative scheme for solving this fixed point equation give rise to the (naive) mean field equations.[12]

The graphical consequence of the naive mean field approximation is to decouple all nodes of the graph. Figure 2.13 illustrates this transformation: the original graph, shown in (a), is a 3×3 grid. The mean field distribution is fully factorized, and so has the

[12] In naive mean field, a fully factorized binary distribution of a binary vector $\mathbf{x} \in \{0,1\}^N$ is represented as $q(\mathbf{x}) = \prod_{s=1}^N \mu_s^{x_s}(1-\mu_s)^{1-x_s}$, where the quantities $\{\mu_s\}$ are the mean field parameters. Taking gradients with respect to μ yields the usual mean field equations. Our exponential parameterization is related via $\theta_s = \log[\mu_s/(1-\mu_s)]$.

structure of the fully disconnected graph shown in (b). The mean field approximation introduces the additional variational parameters μ_s (or $\theta_s = \log[\mu_s/(1-\mu_s)]$), which can be viewed as adjustable inputs to each node. This is typical of a variational transformation: it simplifies the problem (i.e., removes edges) with the additional expense of introducing variational parameters to optimize.

The naive mean field approximation can be quite accurate in certain cases. An important example is a large and densely connected graphical model in which the pairwise couplings between variables are relatively weak. By the law of large numbers, the confluence of many effects on a given node converges to a "mean effect", so that the actual distribution is close to fully factorized. (See Jaakkola [93] for further discussion of such issues.) However, a fully factorized approximation is unable to capture multimodal behavior, and can often be a very poor approximation. For this reason, it is natural to use approximations with more structure, but that nonetheless remain tractable. Natural examples include factorized distributions formed by clustered nodes, as well as tree-structured distributions. Accordingly, different choices of $\mathcal{F}$ — corresponding to distributions with more structure than a fully factorized distribution — lead to more advanced mean field methods. For example, given a particular tree embedded within the original graph with edge set $E_{\text{tree}} \subset E$, we can set

$$\mathcal{F}_{\text{tree}} = \{\ \theta \mid \theta_{st} = 0 \ \ \forall\, (s,t) \ \notin \ E_{\text{tree}}\ \} \tag{2.48}$$

which corresponds to the e-flat manifold of distributions structured according to the tree. This general idea of obtaining approximations richer than a fully factorized distribution is known as *structured mean field*; such approaches were pioneered by Saul and Jordan [155], and have been investigated by a number of other researchers [e.g., 12, 75, 92, 96, 179].

■ 2.3.2 Stationarity conditions for mean field

As noted earlier, mean field optimization, as formulated in equation (2.46), does not fall within the purview of standard information geometry. In particular, although the family $\mathcal{F}$ is an e-flat manifold, the minimization does *not* take place over the second argument (which would correspond to an m-projection), but rather over the first argument. For this reason, mean field theory fails to share the geometry and optimality conditions of the m- or e-projections described in Section 2.2.7. For instance, solutions of mean field equations are not necessarily unique, and can exhibit undesirable properties such as "spontaneous symmetry breaking", in which the mean field solution is asymmetric despite complete symmetry of the actual distribution. See [93] for a simple but compelling example. Nonetheless, mean field solutions do have certain geometric properties, which we develop in this subsection for future reference.

We now derive an alternative set of stationary conditions for mean field in the general case. We first take derivatives of the KL divergence with respect to θ to obtain $\nabla_\theta D(\theta \,\|\, \theta^*) = G(\theta)\,[\theta - \theta^*]$ where $[G(\theta)]_{\alpha\beta} = \operatorname{cov}_\theta\{\phi_\alpha,\, \phi_\beta\}$ is the Fisher information matrix evaluated at θ. Let $\mathcal{J}$ be a subset of the potential index set, such that the approximating family has the form $\mathcal{F}_{\mathcal{J}} = \{ \; \theta \mid \theta_\alpha = 0 \quad \forall \; \alpha \; \notin \; \mathcal{J} \; \}$. Then the stationary conditions for a mean field solution $\widehat{\theta}$ are:

$$\Big[G(\widehat{\theta}) \; (\widehat{\theta} - \theta^*) \Big]_\alpha = \sum_\beta [G(\widehat{\theta})]_{\alpha\beta} \; [\widehat{\theta} - \theta^*]_\beta \; = \; 0 \qquad \forall \, \alpha \, \in \, \mathcal{J} \tag{2.49}$$

From equation (2.29), recall the role of the Fisher information matrix as the Jacobian of the mapping between θ and the dual variables $\eta_\alpha = \mathbb{E}_\theta[\phi_\alpha]$. By a Taylor series expansion, this ensures that

$$[\widehat{\eta} - \eta^*]_\alpha \approx 0 \qquad \forall \; \alpha \; \in \mathcal{J}$$

where the approximate equality holds up to first order in the perturbation $[\widehat{\theta} - \theta^*]$. That is, the mean field stationary conditions in (2.49) ensure that the dual variables $\widehat{\eta}_\alpha$ match the desired statistics η^*_α up to first order for all free indices (i.e., $\alpha \in \mathcal{J}$).

As a concrete illustration, in the case of naive mean field for an Ising model, the mean field stationarity conditions guarantee that

$$\mathbb{E}_{\widehat{\theta}}[x_s] \; = \; p(x_s = 1; \widehat{\theta}) \;\; \approx \;\; p(x_s = 1; \theta^*) \; = \; \mathbb{E}_{\theta^*}[x_s]$$

for all nodes $s \in V$. That is, the single node marginals of the mean field approximation are approximately equal (up to first order) to those of the original model. To emphasize the difference with standard information geometry, recall from Example 2.2.10 that the m-projection of θ^* onto the set of fully factorized distributions $\mathcal{F}_0$ (i.e., computing $\arg\min_{\theta \in \mathcal{F}_0} D(\theta^* \parallel \theta)$) would guarantee the *equality* of these first order marginals.